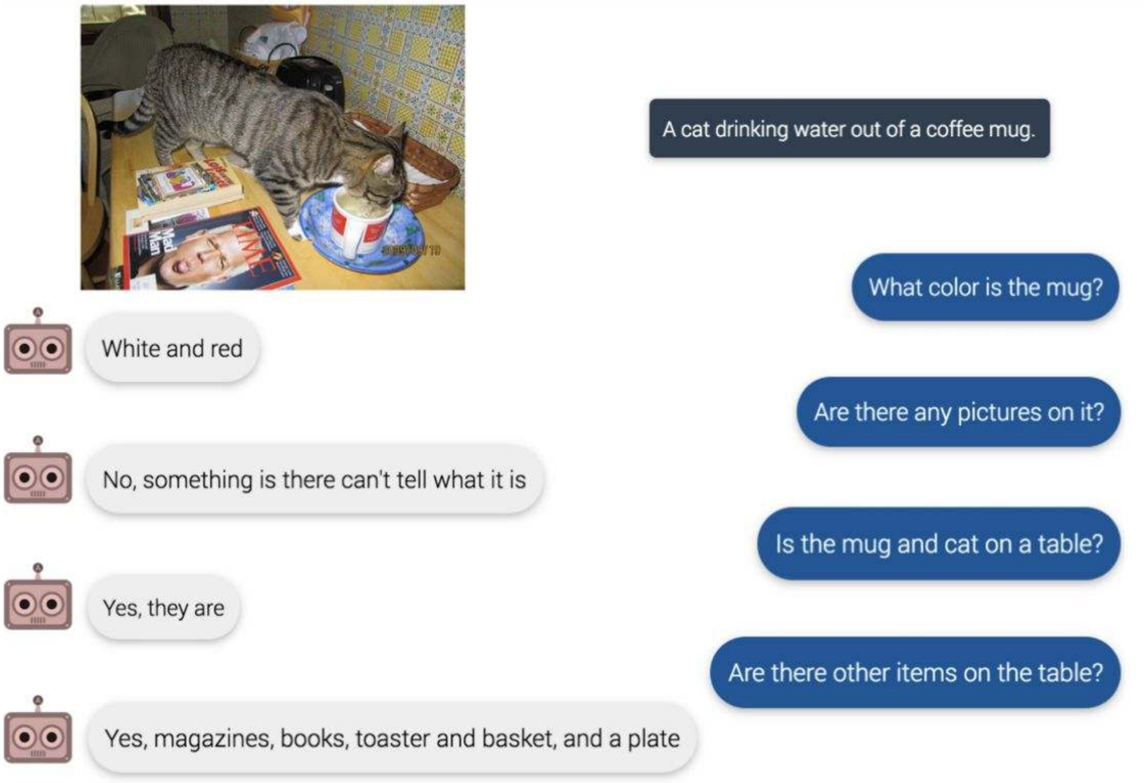
Visual Reference Resolution using Attention Memory for Visual Dialog

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, Leonid Sigal
NIPS 2017

Presented by,
Siddhesh
Anand

Visual Dialog

Task that requires an AI agent to hold a conversation about visual content

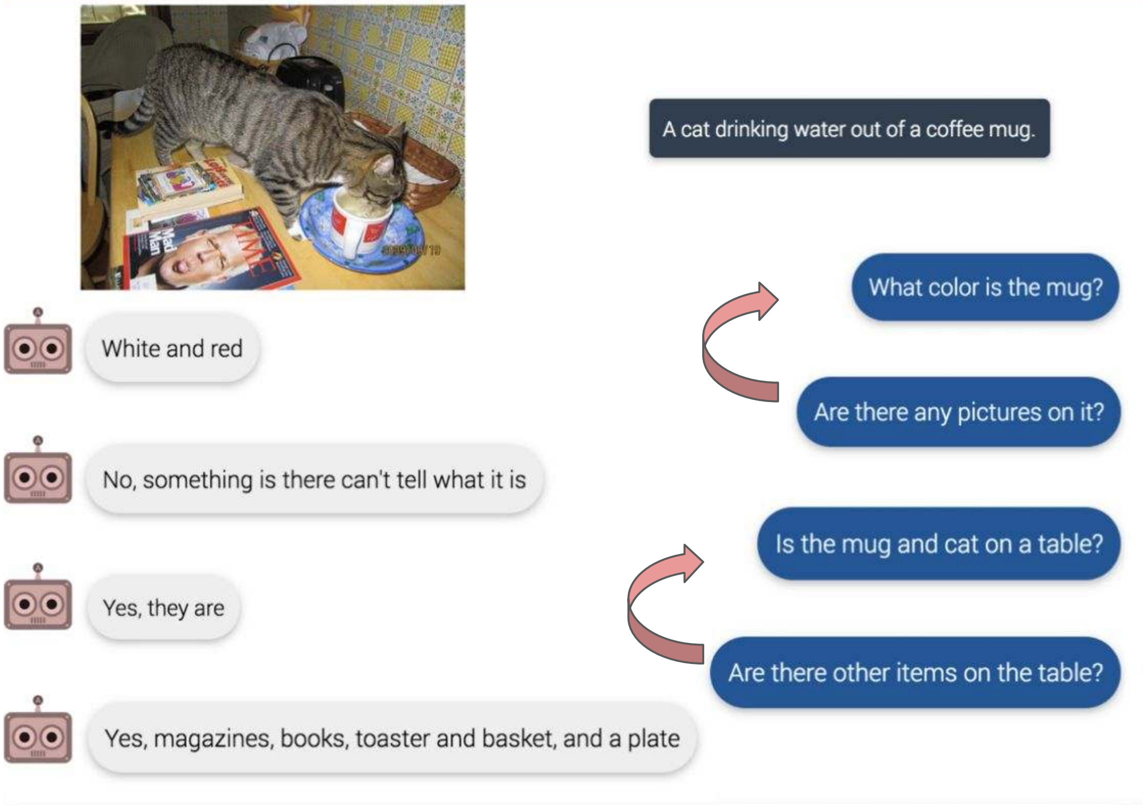


https://visualdialog.org/

Why is this any different from Visual Question Answering?

Key Challenge

Unlike VQA, where every question is asked independently, a visual dialog system needs to answer a sequence of **inter-dependent questions** about an input image.



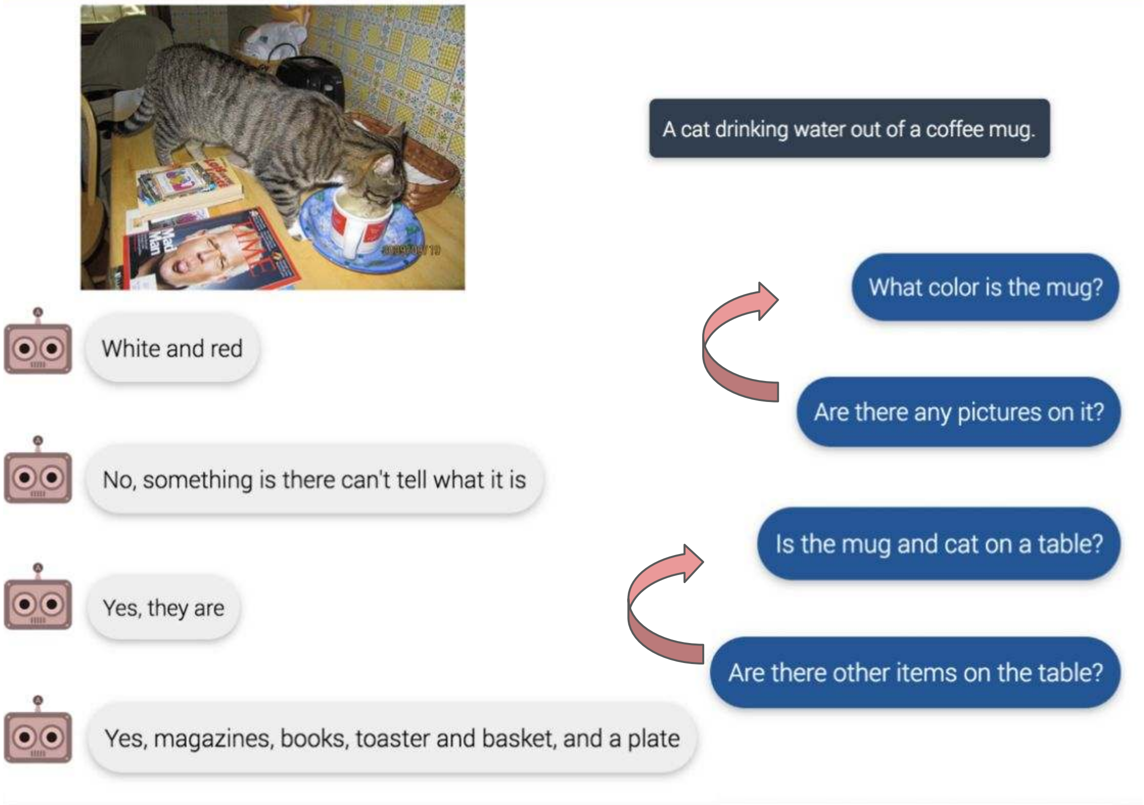
https://visualdialog.org/

Key Challenge

Inter-dependent questions

↓

Ambiguity in Questions

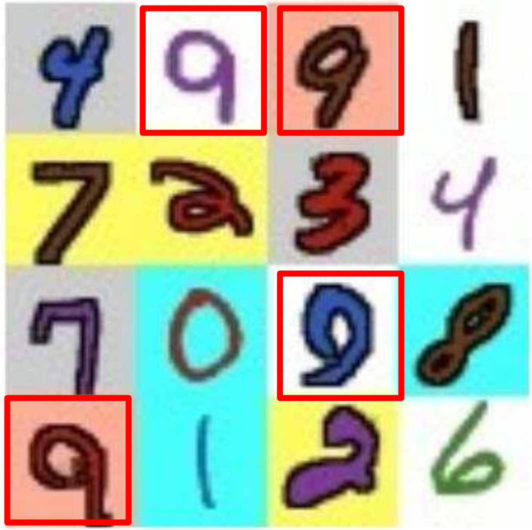


https://visualdialog.org/

Visual Reference Resolution

#	Question	Answer
1	How many 9's are there in the image?	four

Visual Reference Resolution



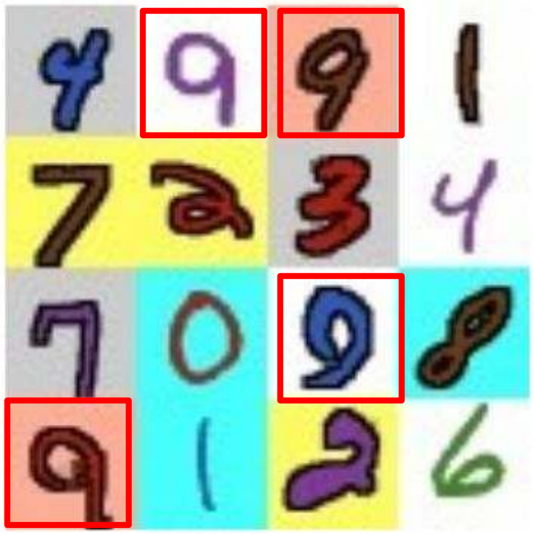
#	Question	Answer
1	How many 9's are there in the image?	four

This can be answered by directly looking at the right regions of the image

Visual Reference Resolution

#	Question	Answer
1	How many 9’s are there in the image?	four
2	How many brown digits are there among them?	one

Visual Reference Resolution



#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

Visual Reference Resolution

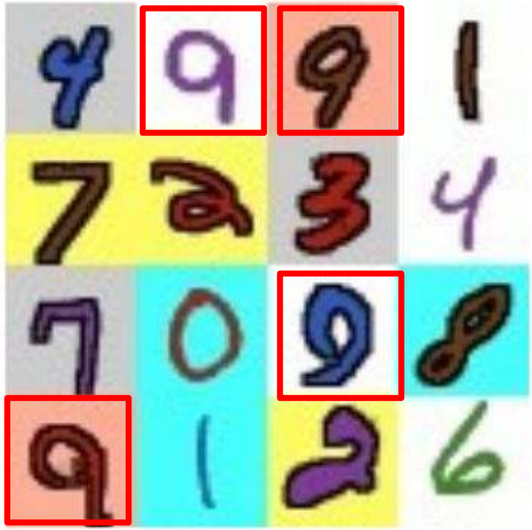
#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one
3	What is the background color of the digit at the left of it?	white
4	What is the style of the digit?	flat
5	What is the color of the digit at the left of it?	blue

Visual Reference Resolution is required to localize attention accurately in the presence of ambiguous expressions

Problem Statement

Improve Visual Dialog by performing Visual Reference Resolution to remove ambiguity in questions

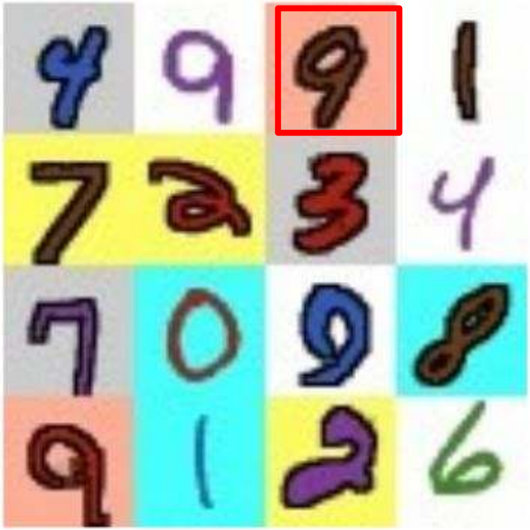
Visual Reference Resolution



#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

To understand what **them** refers to, we need to look at the previous question

Visual Reference Resolution



#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among $\boxed{\text{them?}}$	one

Using information from the previous region of importance is useful!

Visual Reference Resolution

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

Using information from the previous region of importance is useful!

Keep track of all the **previous regions of interest** using an **attention memory**

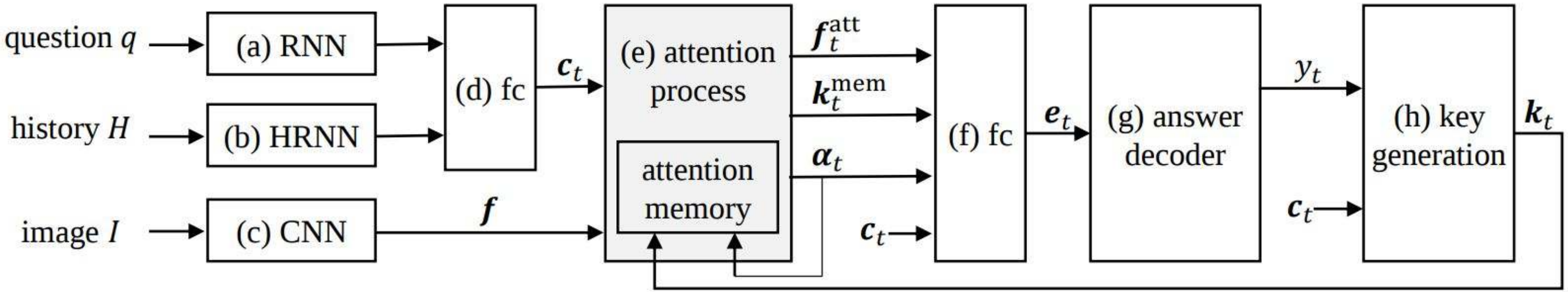
Contributions

- Novel Attention Mechanism to resolve Visual References
    - Use **Associate Attention Memory** to keep track of previous regions of importance in the image
- Comprehensive Analysis of the capacity of the model
- State of the art performance on benchmark datasets

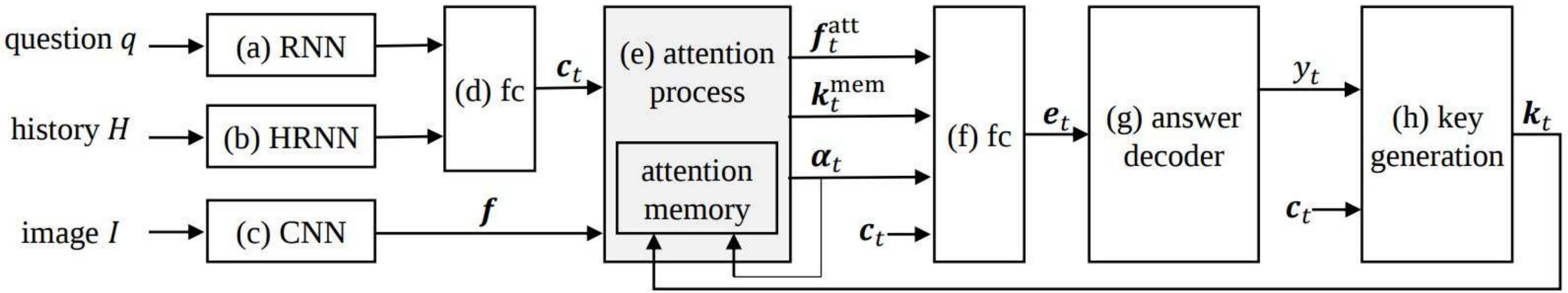
Related Work

- **Visual Dialog**, Das et al., CVPR 2017
    - Used **Memory** to actively select the previous question in the history
- **Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning**, Das et al., ICCV 2017
    - Deep RL based approach
- **GuessWhat?! Visual object discovery through multi-modal dialogue,** Vries et al., CVPR 2017
    - Object discovery through a series of yes/no questions

Model Overview



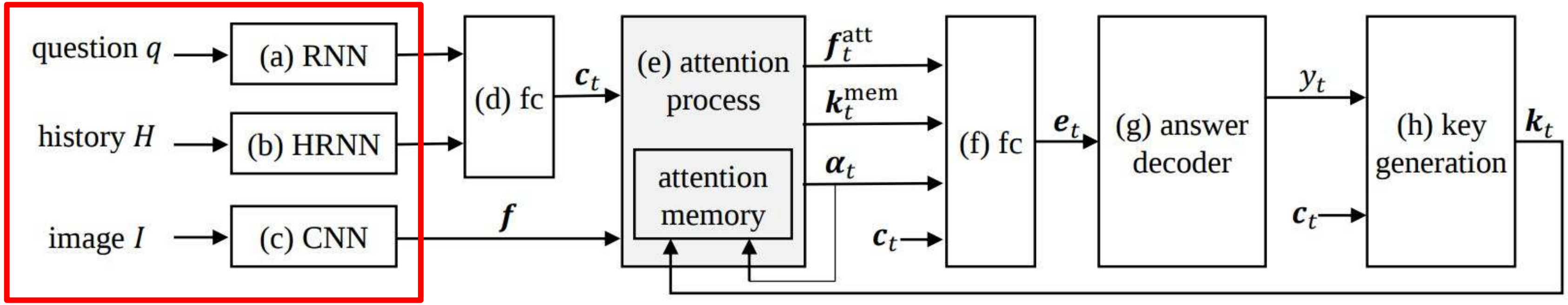
Task



Predict an answer $\mathbf{y}_t$ at time t given

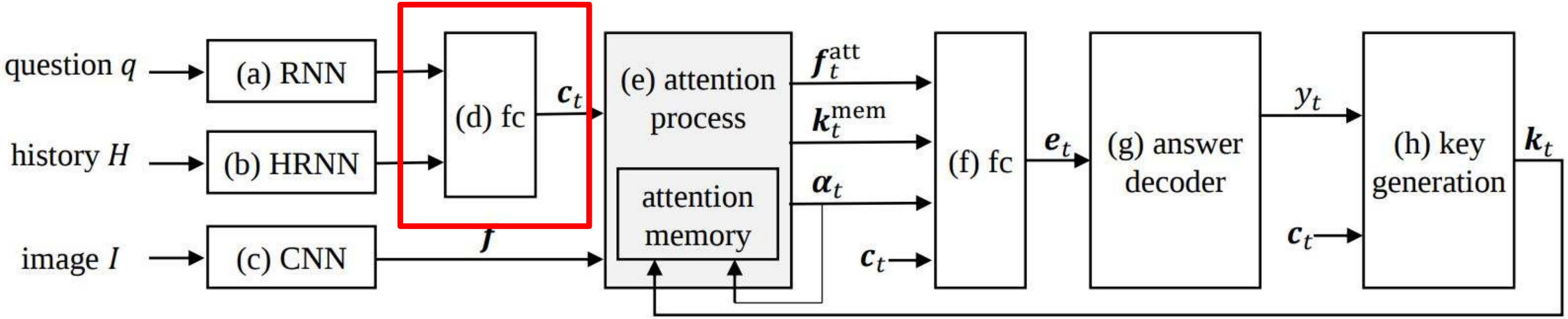
- input image **I**,
- current question $\mathbf{q}_t$
- dialog history $\mathbf{H} = \{(q_1, y_1), (q_2, y_2), \ldots, (q_t, y_t)\}$

Encoding



- LSTM to encode question $\mathbf{y}_t$
- CNN to extract to compute a feature map **f**
- Hierarchical LSTM to encode the history
    - Encode the question and answer separately using LSTMs
    - Obtain a QA embedding by passing it through a **fc layer**
    - QA embedding is passed through another LSTM to get **history encoding**

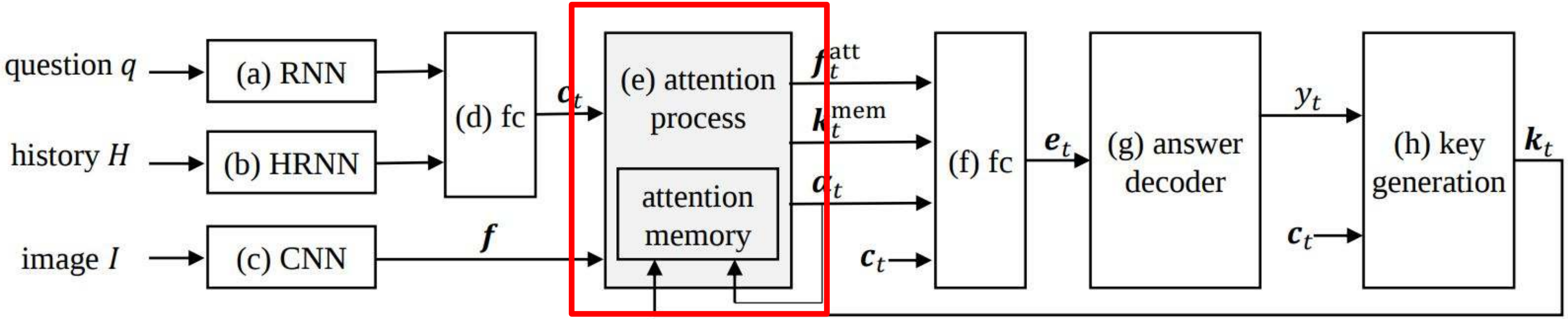
Encoding



- **fc layer** to obtain a fused encoding of the history and question
  - This serves as a context embedding which will be helpful when looking at what region of the image to consider

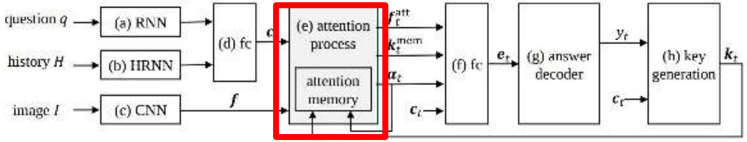
$$c_t = \mathtt{fc}(\mathrm{RNN}(q_t), \mathrm{HRNN}(H))$$

Attention



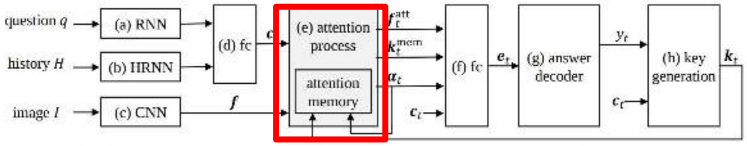
- **Two** types of attention that focus on different aspects
  - **Tentative Attention**: How important is the current question
  - **Associative Attention Memory**: How important are the previous questions

Tentative Attention

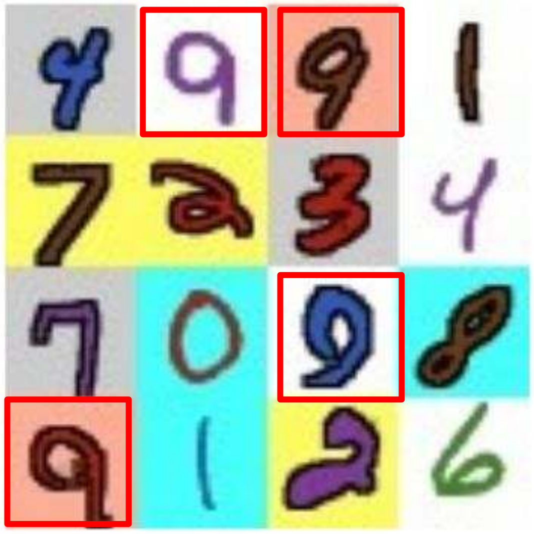


- **Key Idea:** For certain questions, we can directly look at the image to figure out the answer

Tentative Attention

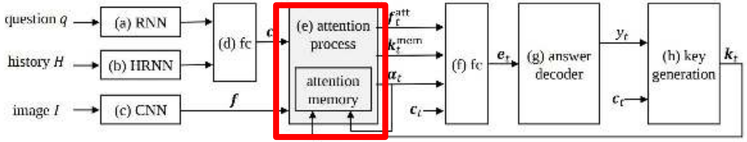


- **Key Idea:** For certain questions, we can directly look at the image to figure out the answer



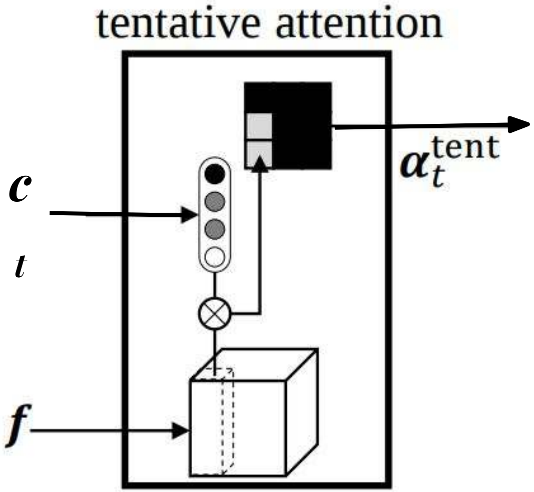
#	Question	Answer
1	How many 9's are there in the image?	four

Tentative Attention

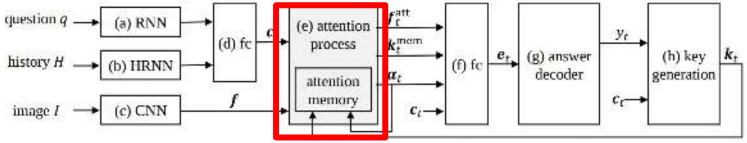


- **Key Idea:** For certain questions, we can directly look at the image to figure out the answer

$$s_{t,n} = \left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)^\top \left(\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n\right)$$

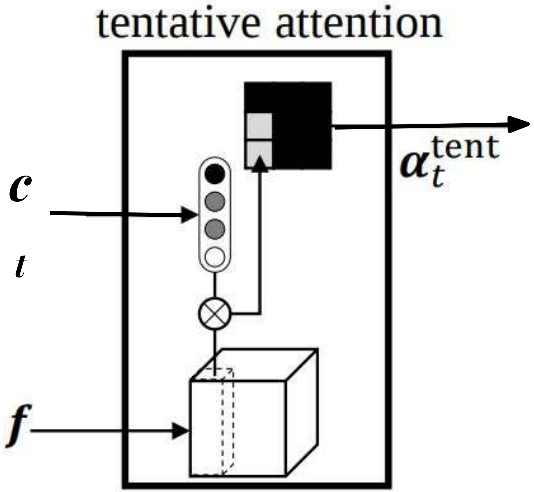


Tentative Attention

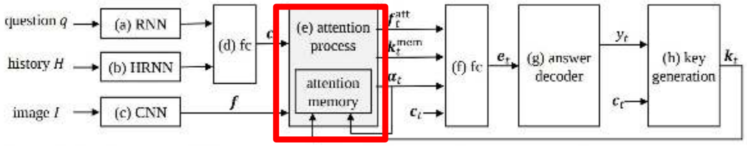


- **Key Idea:** For certain questions, we can directly look at the image to figure out the answer

$$s_{t,n} = \overbrace{(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t)}^{\text{Projection of context vector into some space}}{}^{\top} \, (\underbrace{\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n}_{\text{Projection of image features vector into the same space}})$$



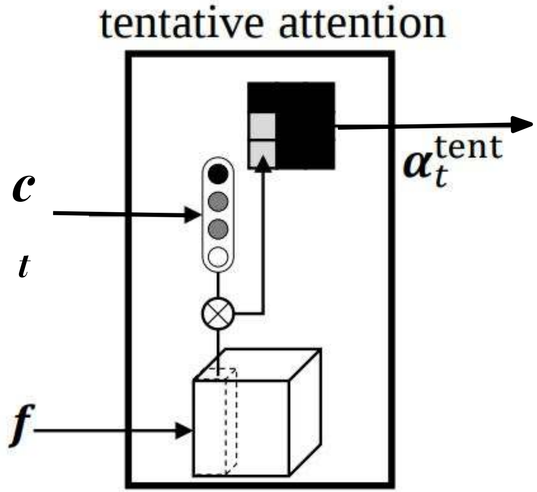
Tentative Attention



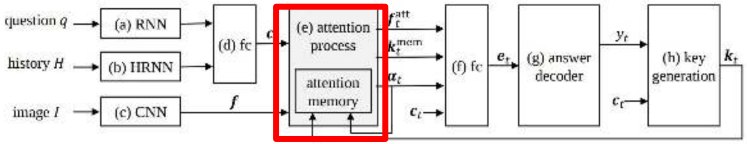
- **Key Idea:** For certain questions, we can directly look at the image to figure out the answer

$$s_{t,n} = \overbrace{\left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)}^{\text{Projection of context vector into some space}}{}^\top \left(\underbrace{\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n}_{\text{Projection of image features vector into the same space}}\right)$$

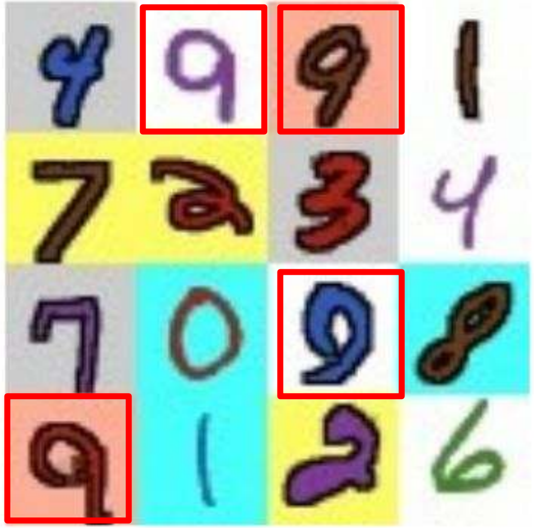
$$\boldsymbol{\alpha}_t^{\text{tent}} = \text{softmax}\left(\{s_{t,n}, 1 < n < N\}\right)$$



Attention Memory

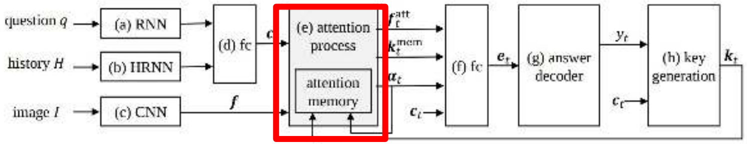


- **Key Idea:** Explicitly store the image attentions obtained in the past

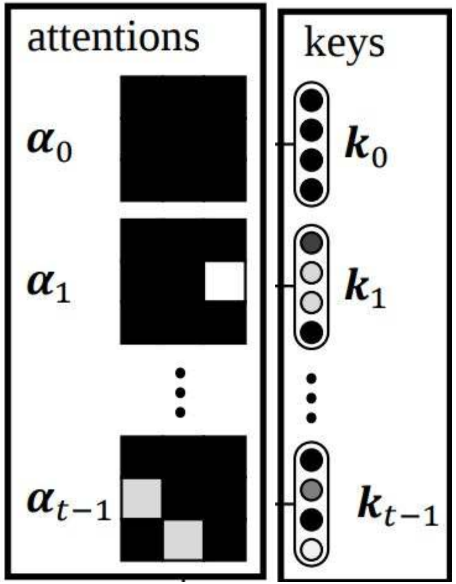


#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

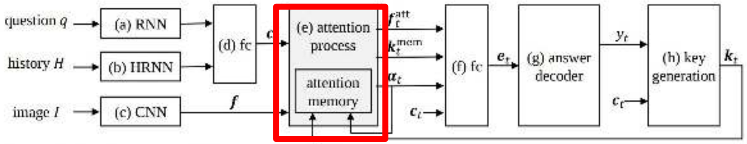
Attention Memory



- **Key Idea:** Explicitly store the image attentions obtained in the past
- Every item in the memory is a **(attention, key)** pair
    - $\boldsymbol{\alpha}_t$ is the attention map at time **t**
    - **k** is the key which captures the dialog history (including answers) so far

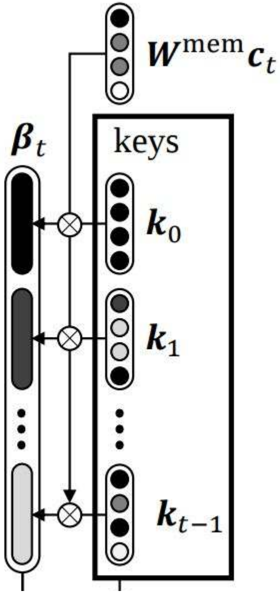


Attention Memory

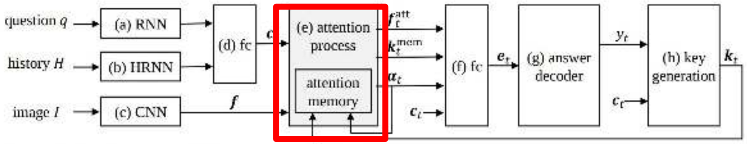


- **Key Idea:** Explicitly store the image attentions obtained in the past
- Every item in the memory is a **(attention, key)** pair
  - $\boldsymbol{\alpha}_t$ is the attention map at time **t**
  - **k** is the key which captures the dialog history (including answers) so far

$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t\right)^{\top} \boldsymbol{k}_\tau$$

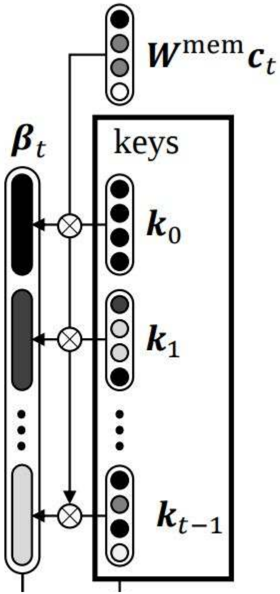


Attention Memory

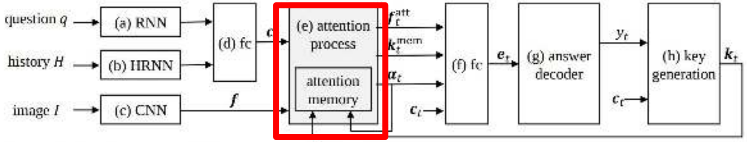


- **Key Idea:** Explicitly store the image attentions obtained in the past
- Every item in the memory is a **(attention, key)** pair
  - $\boldsymbol{\alpha}_t$ is the attention map at time **t**
  - **k** is the key which captures the dialog history (including answers) so far

$$m_{t,\tau} = (\underbrace{\boldsymbol{W}^{\text{mem}}\boldsymbol{c}_t}_{\text{Projection of context vector into some space}})^\top \boldsymbol{k}_\tau$$

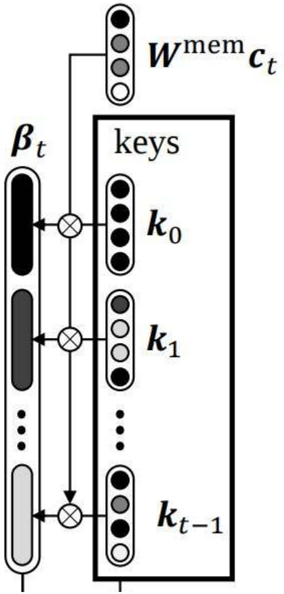


Attention Memory

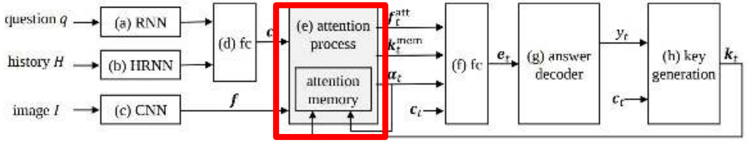


- **Key Idea:** Explicitly store the image attentions obtained in the past
- Every item in the memory is a **(attention, key)** pair
  - $\boldsymbol{\alpha}_t$ is the attention map at time **t**
  - **k** is the key which captures the dialog history (including answers) so far

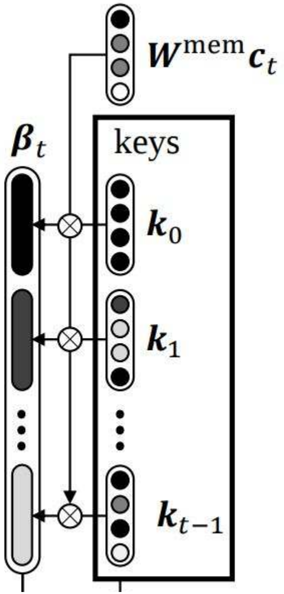
$$m_{t,\tau} = \underbrace{\big(\overbrace{\boldsymbol{W}^{\mathrm{mem}} \boldsymbol{c}_t}^{\text{Projection of context vector into some space}}\big)^{\top} \boldsymbol{k}_\tau}_{\text{Similarity with each key}}$$



Attention Memory

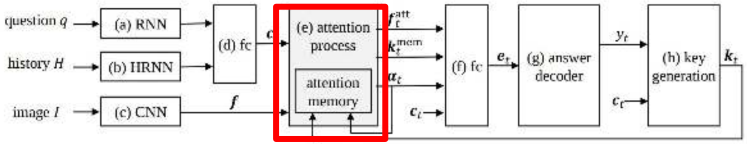


$$m_{t,\tau} = \underbrace{(\overbrace{\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t}^{\text{Projection of context vector into some space}})^\top \boldsymbol{k}_\tau}_{\text{Similarity with each key}}$$



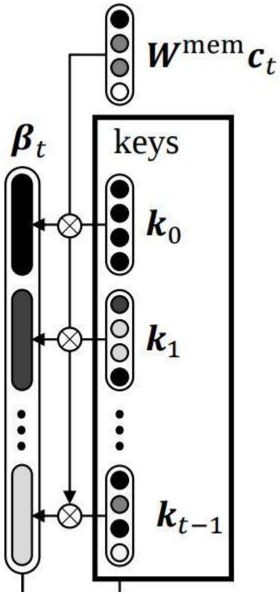
- **Intuition:** How similar is my current context to each of the previous responses

Attention Memory

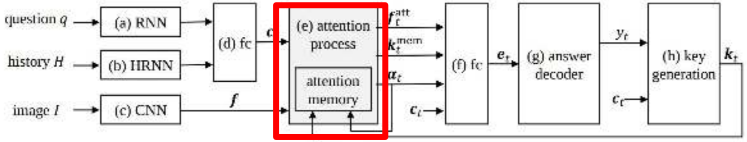


$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$



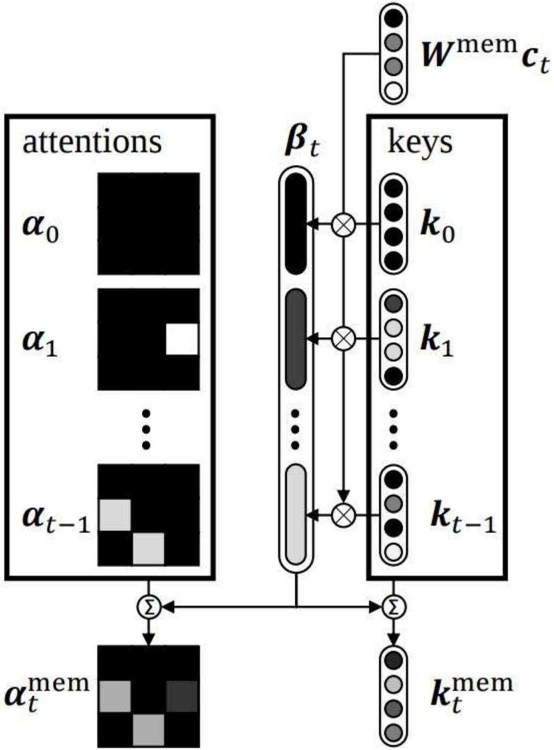
Attention Memory



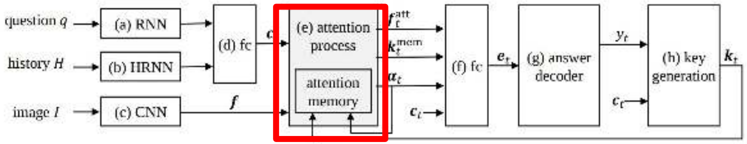
$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t\right)^{\top} \boldsymbol{k}_{\tau}$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$

$$\boldsymbol{\alpha}_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \boldsymbol{\beta}_{t,\tau} \boldsymbol{\alpha}_{\tau} \qquad \boldsymbol{k}_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \boldsymbol{\beta}_{t,\tau} \boldsymbol{k}_{\tau}$$



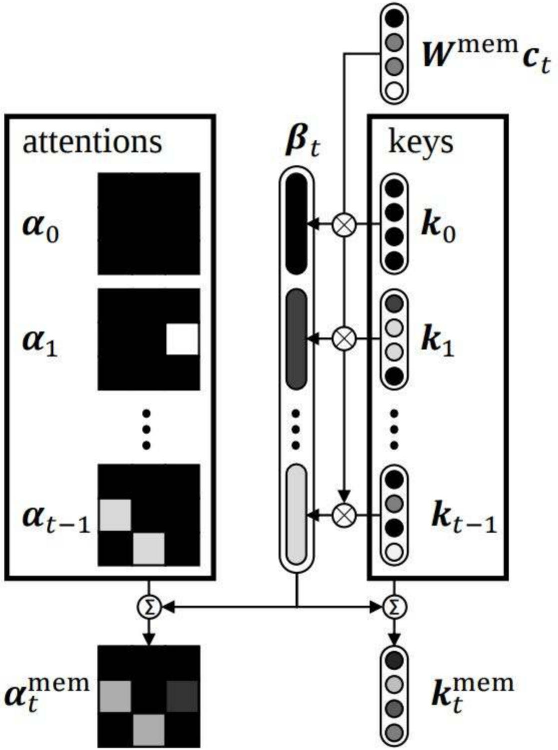
Attention Memory



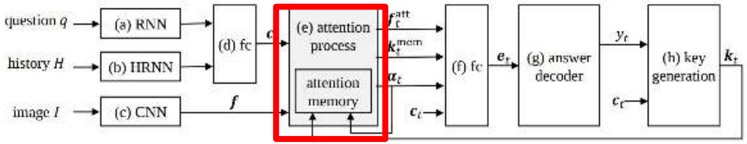
$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}}\boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$

$$\boldsymbol{\alpha}_t^{\text{mem}} = \underbrace{\sum_{\tau=0}^{t-1} \boldsymbol{\beta}_{t,\tau}\boldsymbol{\alpha}_\tau}_{\text{Convex combination of attention maps}} \qquad \boldsymbol{k}_t^{\text{mem}} = \underbrace{\sum_{\tau=0}^{t-1} \boldsymbol{\beta}_{t,\tau}\boldsymbol{k}_\tau}_{\text{Convex combination of keys}}$$

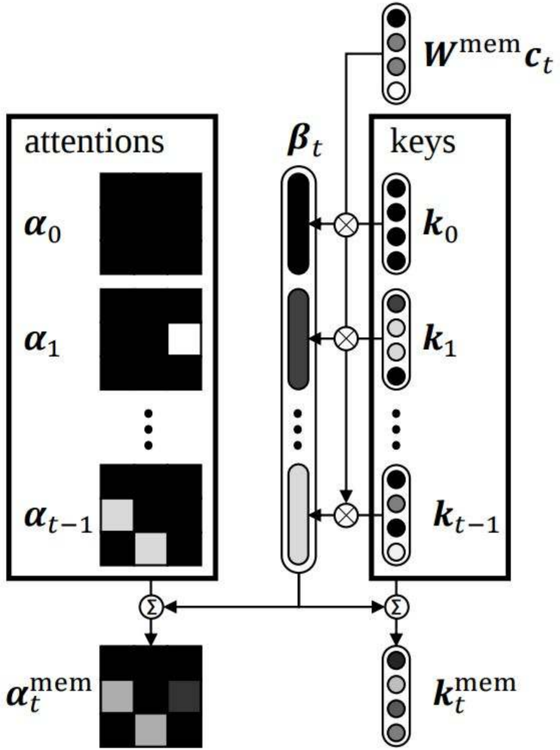


Sequential Dialog

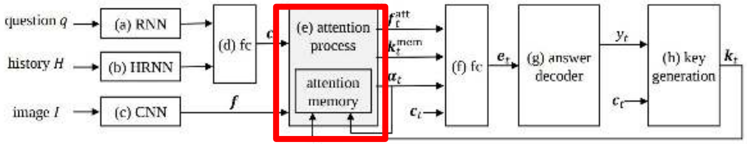


- **Key Idea:** Questions in a dialog have a sequential structure
- Questions that are recent might be more relevant

$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$



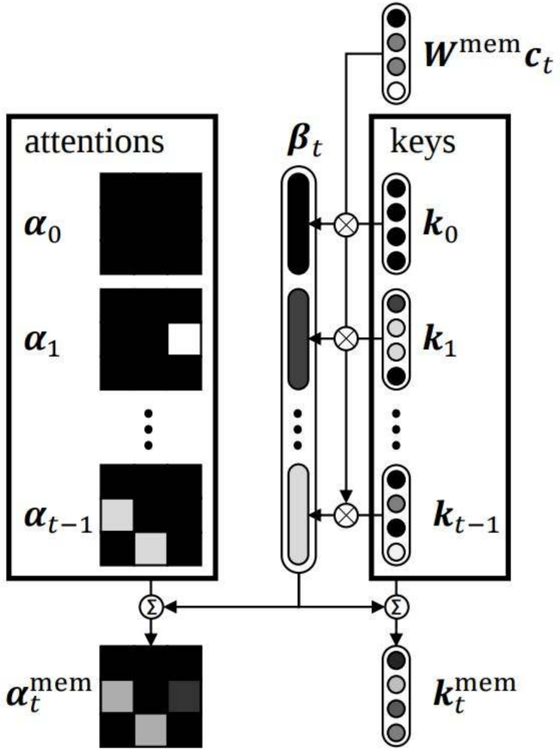
Sequential Dialog



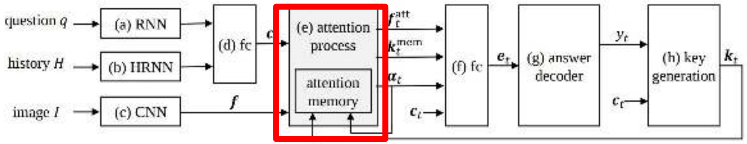
- **Key Idea:** Questions in a dialog have a sequential structure
- Questions that are recent might be more relevant

$$m_{t,\tau} = \left(\boldsymbol{W}^{\mathrm{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$

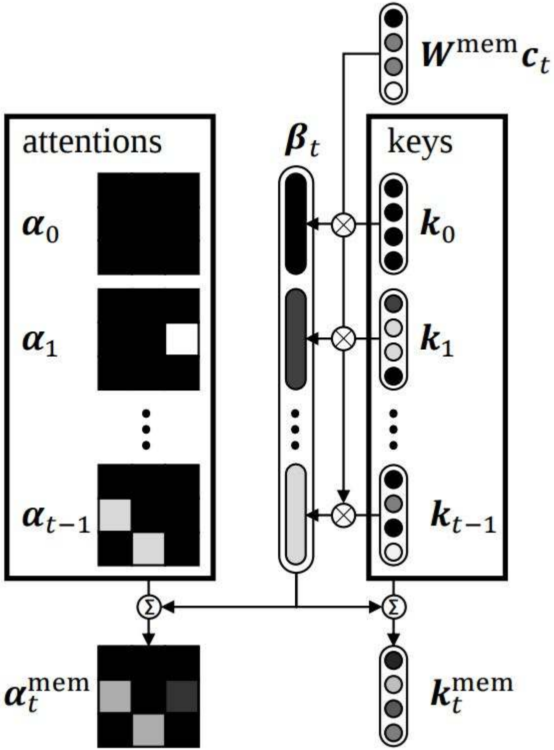
Gives all keys equal weight irrespective of recency



Sequential Dialog



- **Key Idea:** Questions in a dialog have a sequential structure
- Questions that are recent might be more relevant

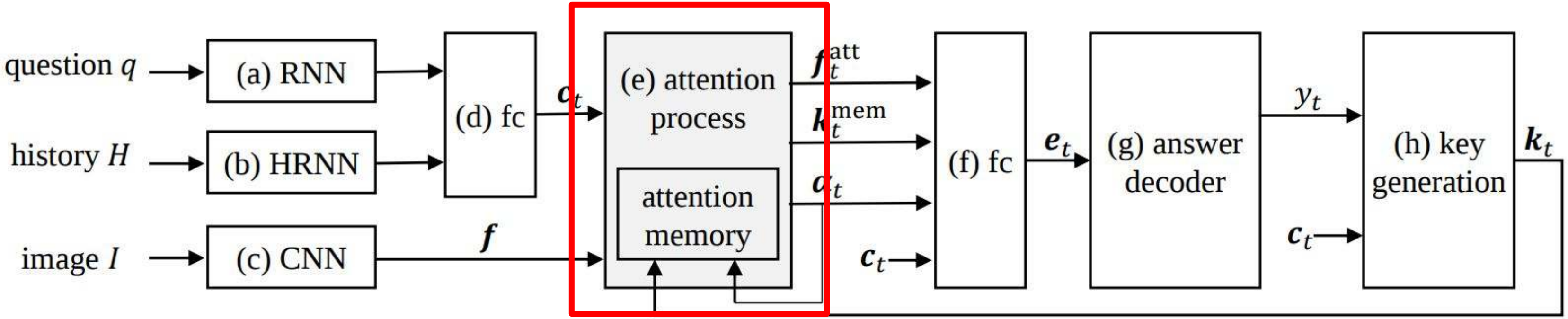


$$m_{t,\tau} = \left(\boldsymbol{W}^{\mathrm{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$

$$m'_{t,\tau} = \left(\boldsymbol{W}^{\mathrm{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau + \theta\,(t - \tau)$$

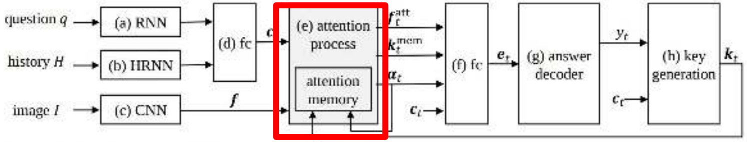
θ is a learnable parameter weighting the relative time distance

Dynamic Attention Combination



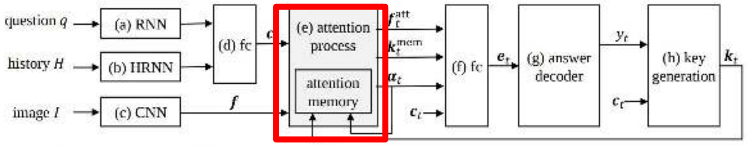
- Combine the **tentative** and **memory attention** to get an attention map for the current question

Dynamic Combination



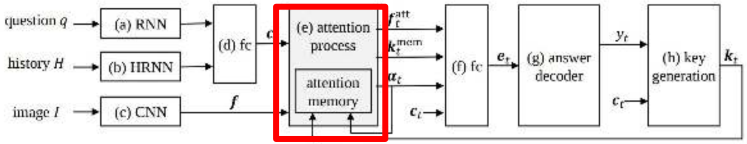
$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\boldsymbol{W}^{\mathrm{DPL}}(\boldsymbol{c}_t) \cdot \underbrace{\gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})}_{\text{Output of a Convolution Layer}}\right)$$

Dynamic Combination



$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\underbrace{\boldsymbol{W}^{\mathrm{DPL}}\left(\boldsymbol{c}_t\right) \cdot \underbrace{\gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})}_{\text{Output of a Convolution Layer}}}_{\text{fc layer}}\right)$$

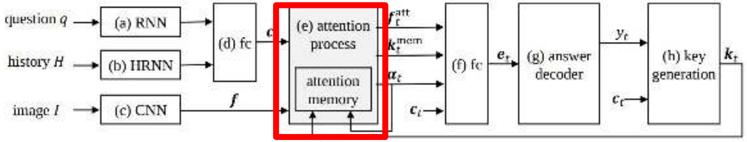
Dynamic Combination



$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\boldsymbol{W}^{\mathrm{DPL}}(\boldsymbol{c}_t) \cdot \gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})\right)$$

If this is simply learned as a parameter, the merging process would not depend on the question

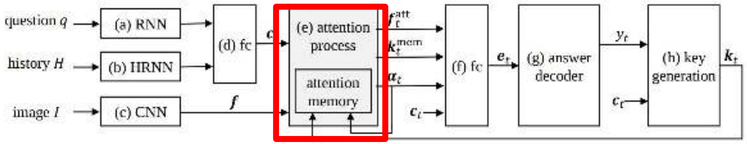
Dynamic Combination

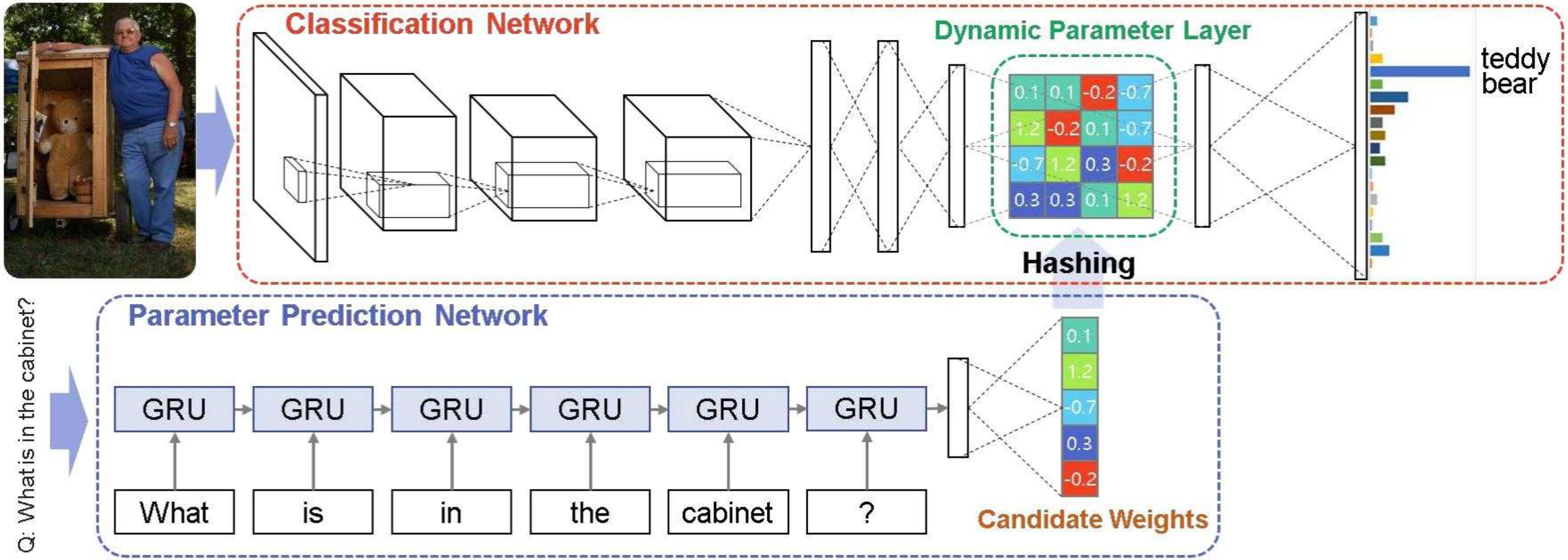


$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\boldsymbol{W}^{\mathrm{DPL}}(\boldsymbol{c}_t) \cdot \gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})\right)$$

- **Key Idea:** Dynamically generate the weight matrix of the fc layer depending on the question

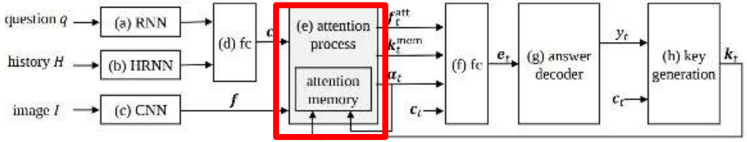
Dynamic Combination





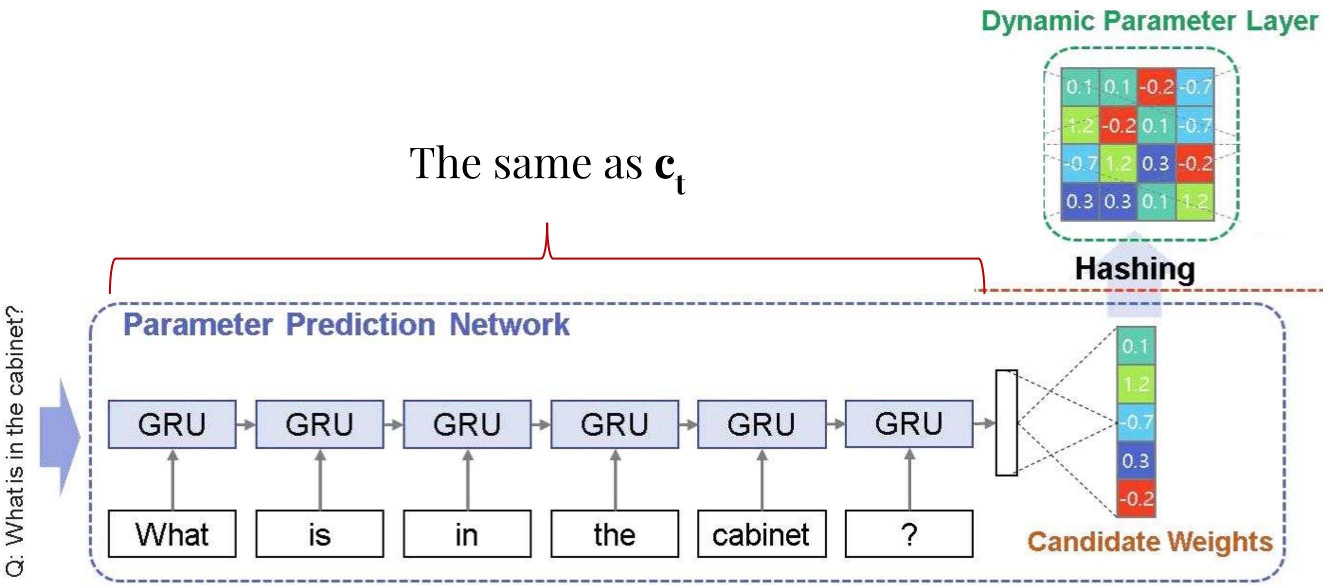
Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Dynamic Combination



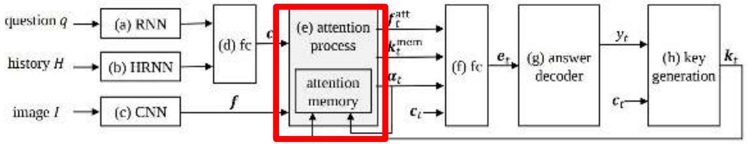
$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \text{softmax}\left(\boldsymbol{W}^{\text{DPL}}(\boldsymbol{c}_t) \cdot \gamma(\boldsymbol{\alpha}_t^{\text{tent}}, \boldsymbol{\alpha}_t^{\text{mem}})\right)$$

The same as $\mathbf{c}_t$



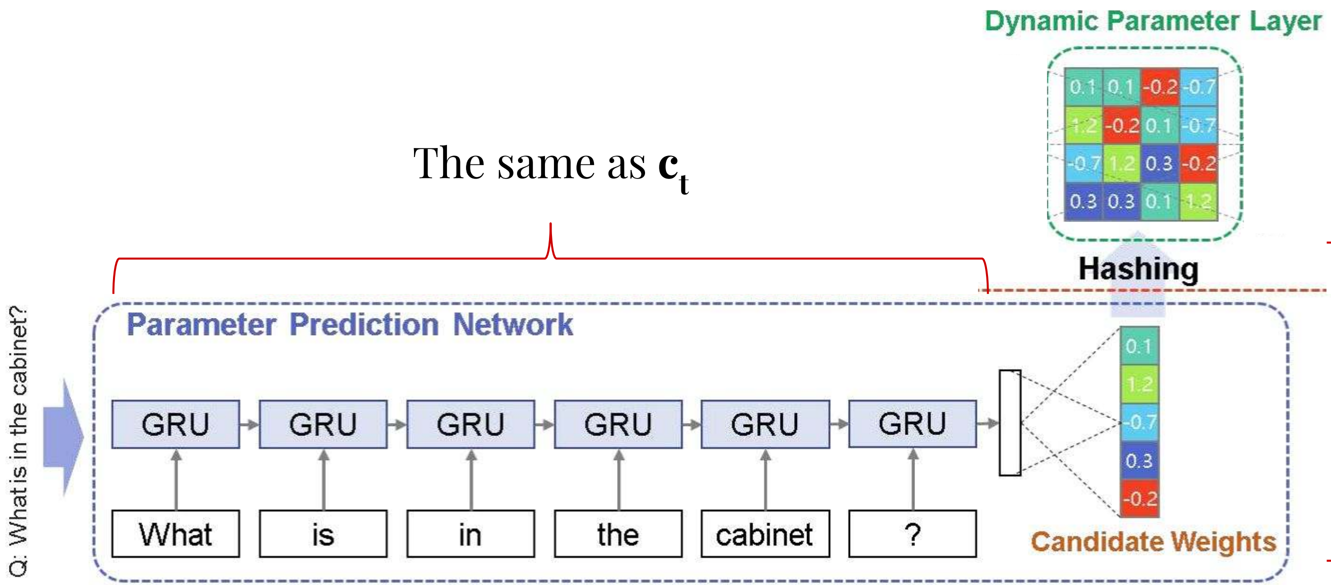
Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Dynamic Combination



$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\boldsymbol{W}^{\mathrm{DPL}}(\boldsymbol{c}_t) \cdot \gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})\right)$$

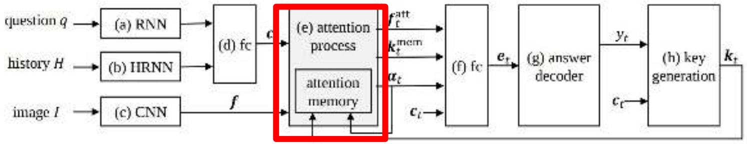
The same as $\mathbf{c}_t$



Hashing trick to reduce complexity

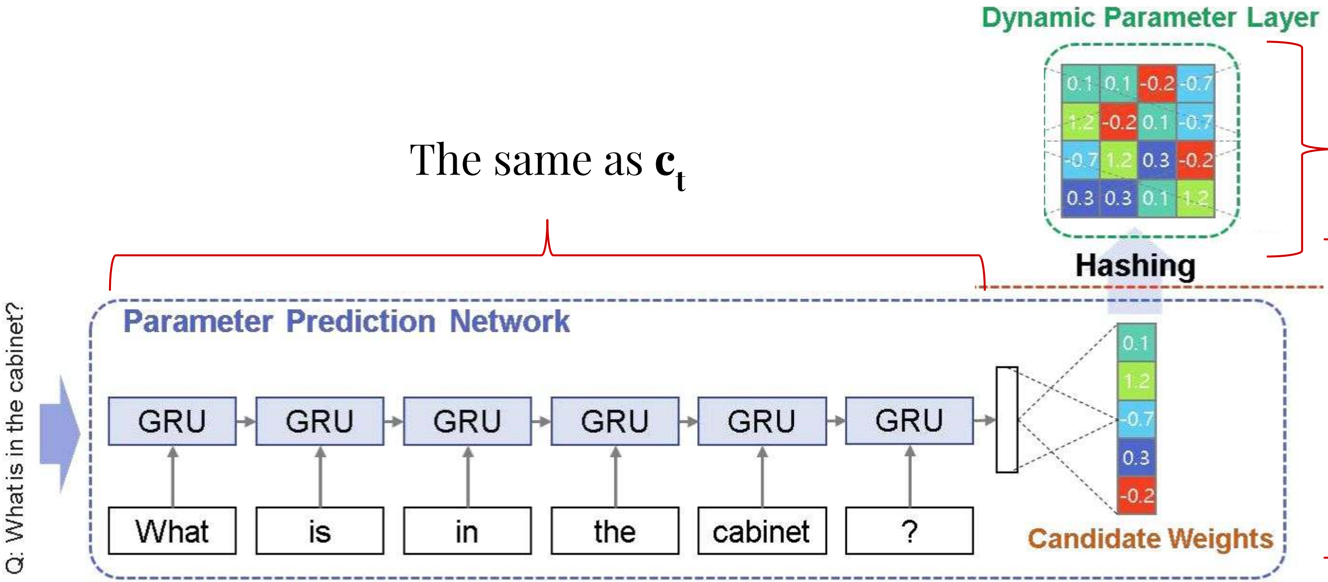
Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Dynamic Combination



$$\boldsymbol{\alpha}_t(\boldsymbol{c}_t) = \mathrm{softmax}\left(\boldsymbol{W}^{\mathrm{DPL}}\left(\boldsymbol{c}_t\right) \cdot \gamma(\boldsymbol{\alpha}_t^{\mathrm{tent}}, \boldsymbol{\alpha}_t^{\mathrm{mem}})\right)$$

The same as $\mathbf{c}_t$

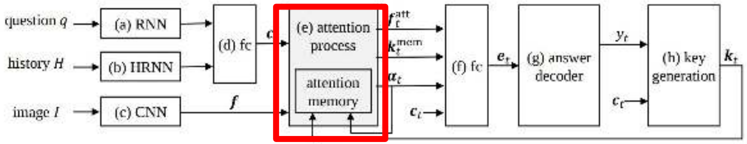


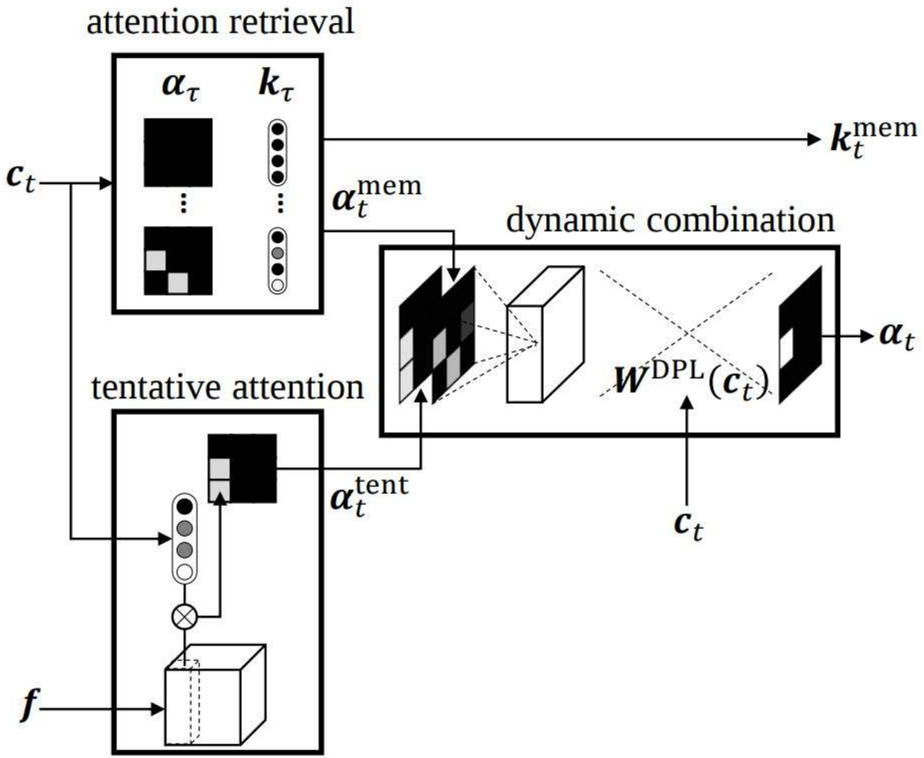
$\boldsymbol{W}^{\mathrm{DPL}}\left(\boldsymbol{c}_t\right)$

Hashing trick to reduce complexity

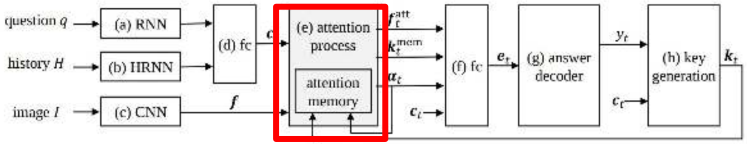
Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

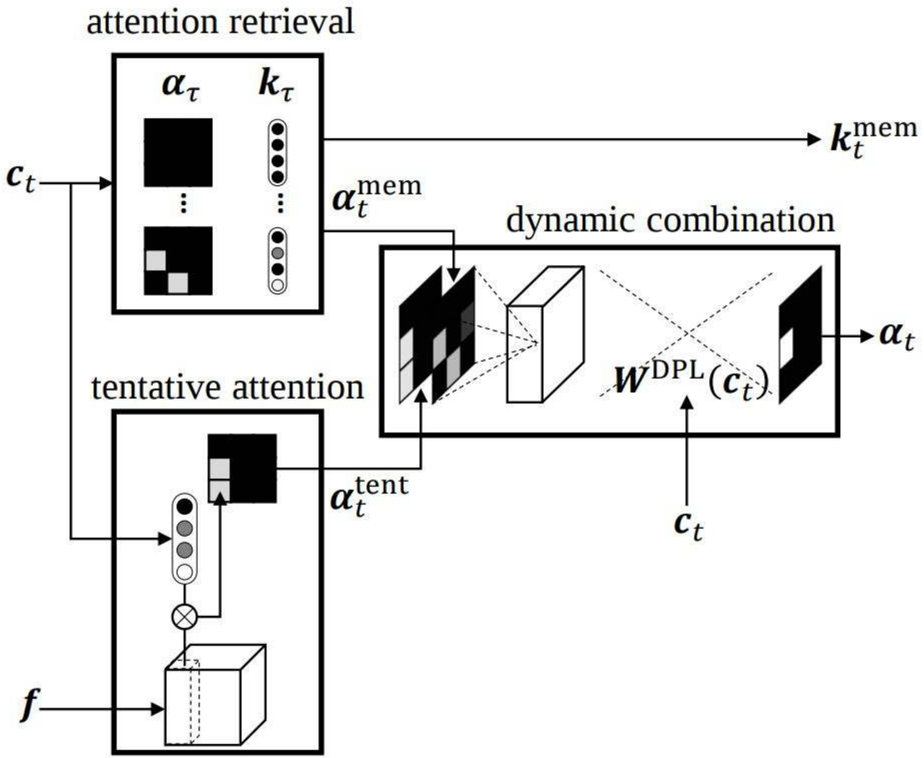
Final Overview





Final Overview

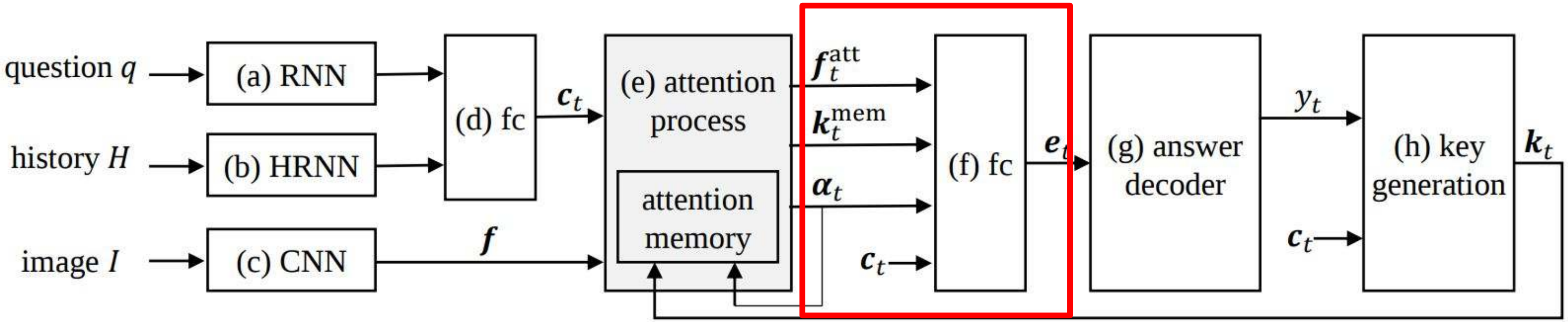




$$f_t^{\text{att}} = [\alpha_t(c_t)]^\top \cdot f$$

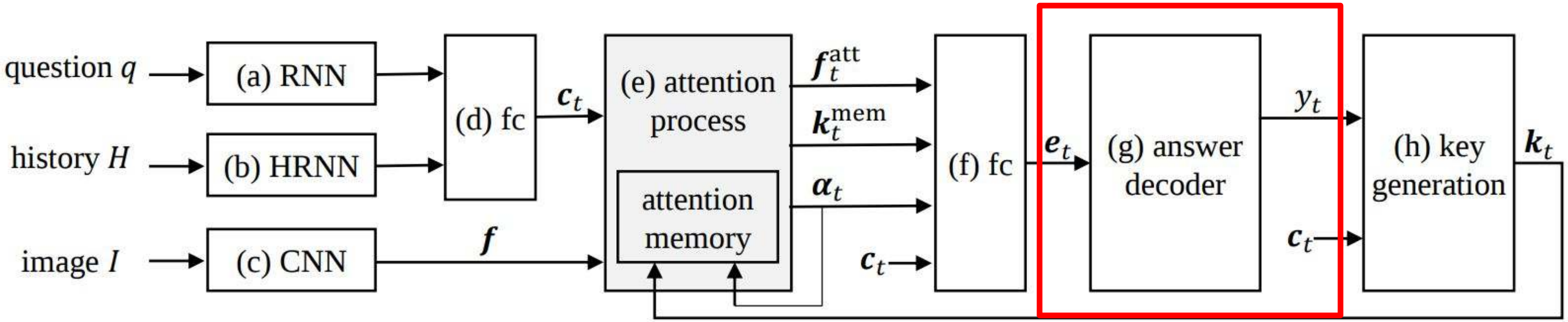
Convex Combination of Image Features

Final Encoding Generation



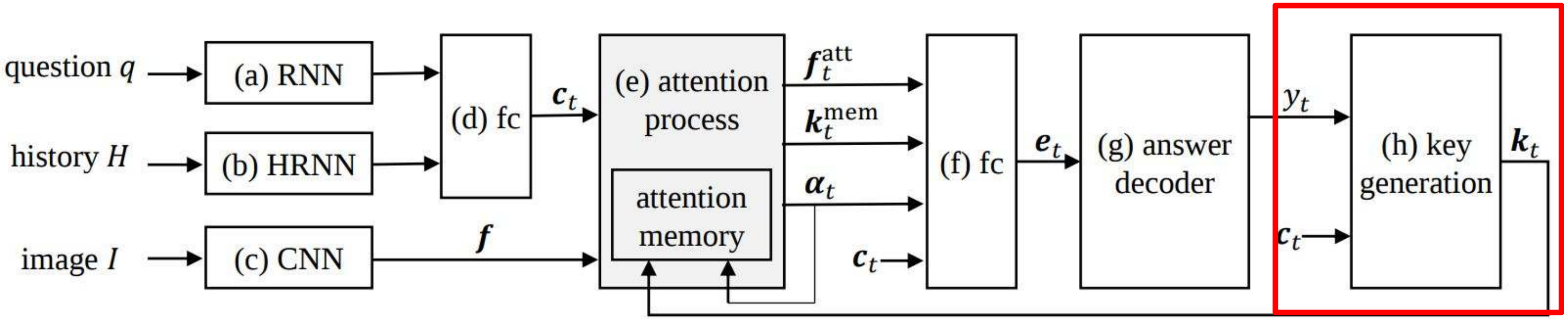
- Combine **attended image feature embedding, context embedding, attention map** and **retrieved key**
  - Intuition is that additional information might be helpful
- e_t is the final encoding

Decoder



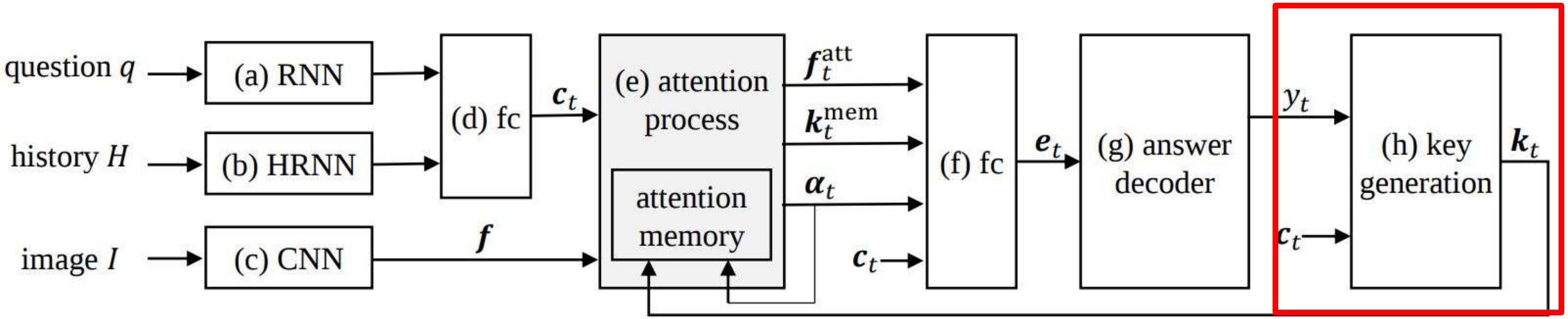
- e_t is used as the hidden state of the LSTM that generates the output y_t

Key Generation



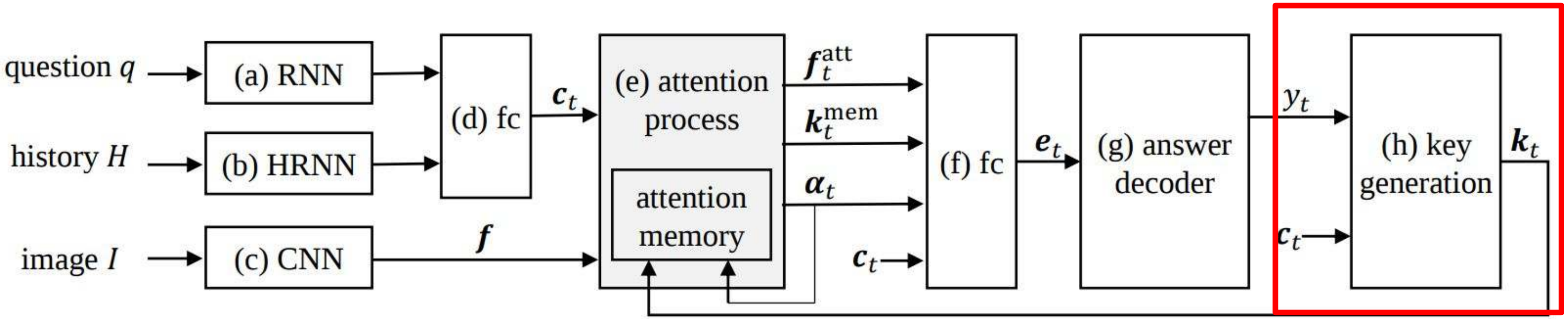
- Answer embedding $\mathbf{a}_t$ is generated by passing $\mathbf{y}_t$ through a LSTM

Key Generation



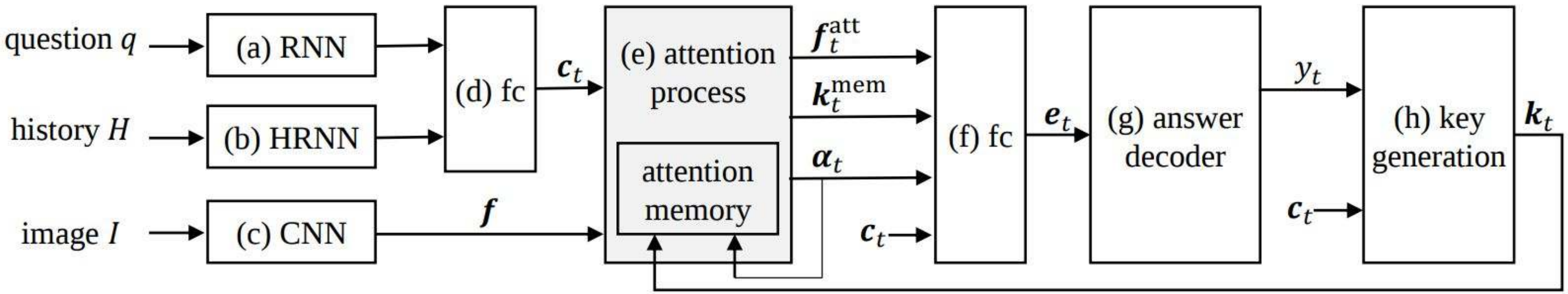
- Answer embedding $\mathbf{a}_t$ is generated by passing $\mathbf{y}_t$ through a LSTM
- Answer embedding $\mathbf{a}_t$ and context embedding $\mathbf{c}_t$ are combined using a **fc layer** to obtain key $\mathbf{k}_t$

Key Generation



- Answer embedding $\mathbf{a_t}$ is generated by passing $\mathbf{y_t}$ through a LSTM
- Answer embedding $\mathbf{a_t}$ and context embedding $\mathbf{c_t}$ are combined using a **fc layer** to obtain key $\mathbf{k_t}$
- The key-attention pair $(\mathbf{k_t}, \boldsymbol{\alpha_t})$ is added to the memory

Training



- Since all the modules of the network are fully differentiable, the entire network can be trained end-to-end by standard gradient-based learning algorithms

Experiments

- MNIST Dialog
  - To measure the ability to resolve visual references
  - Contains ambiguous expressions and strong inter-dependencies
- VisDial
  - Performance in real world dataset

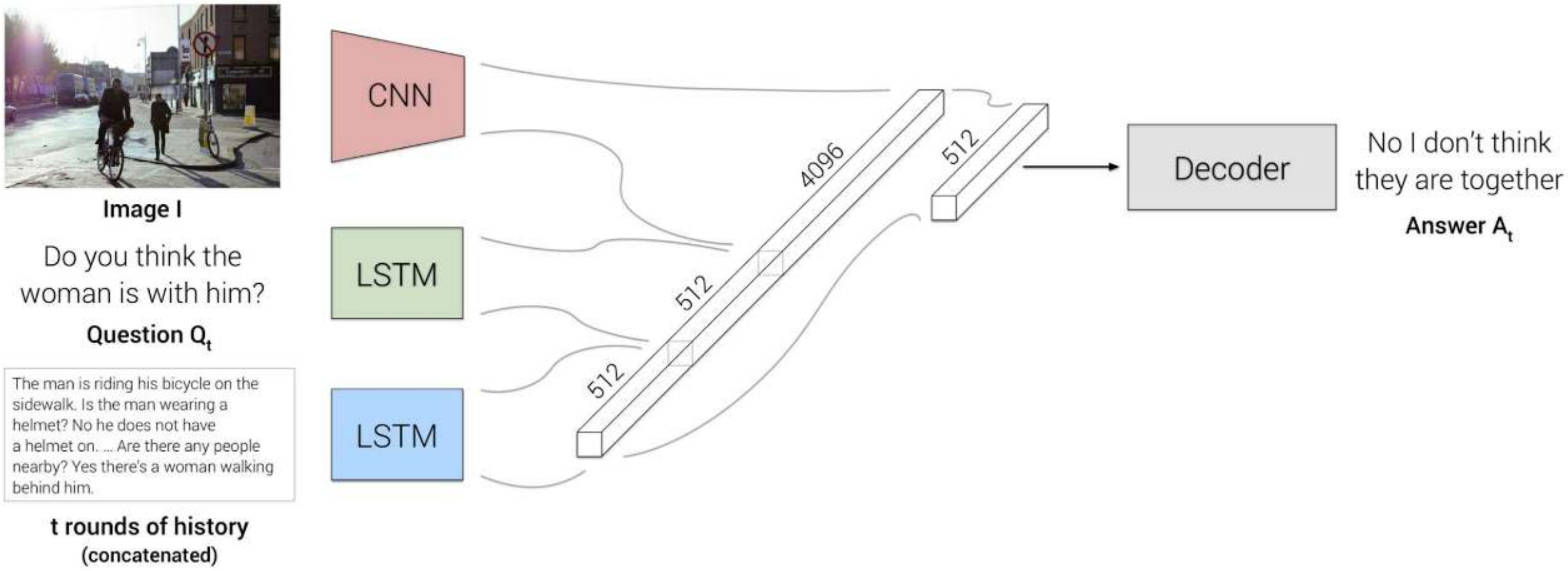
MNIST Dialog

- 4x4 grid of MNIST digits
- Each digit has 4 attributes
  - color(5)
  - bgcolor(5)
  - number(10)
  - style(2)
- Questions
  - Counting
  - Attribute
- Answers
  - Single word

Implementation details

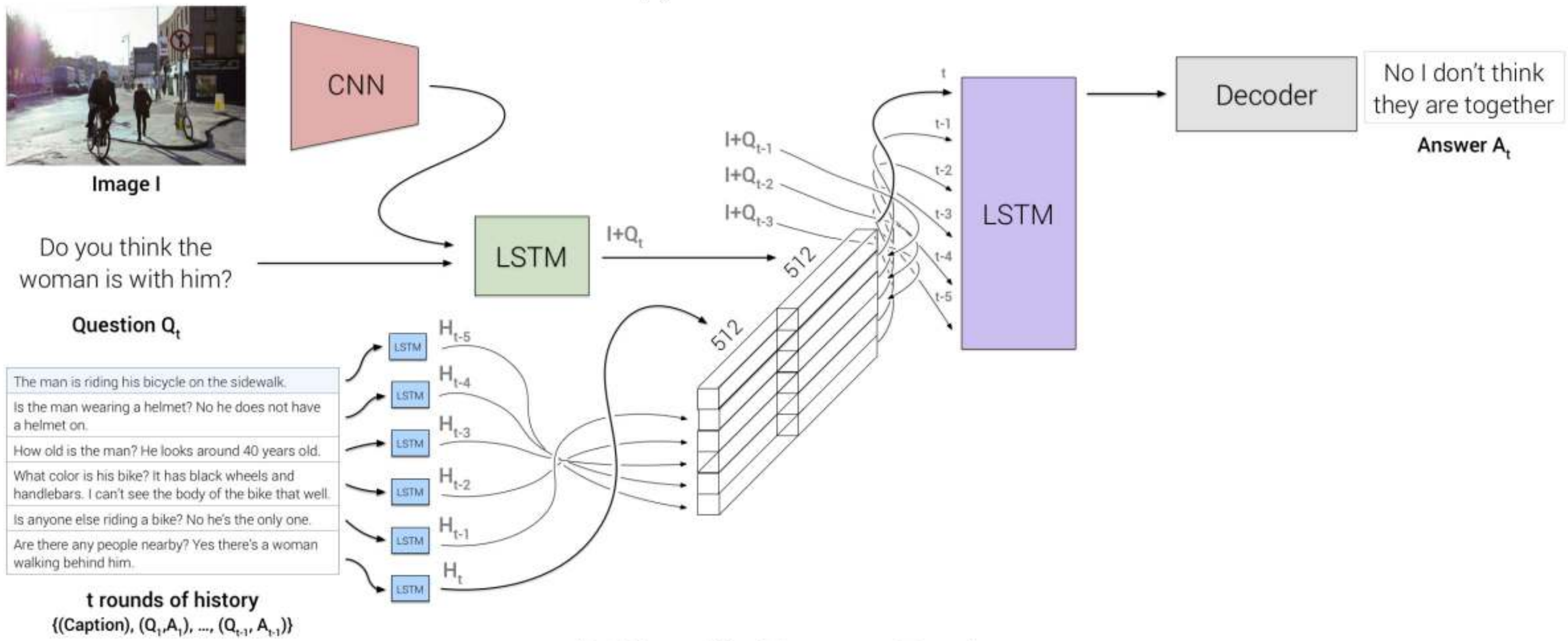
- Word embedding (32 x 1)
- Hidden state (64 x 1)
- 4 Convolution layer (3 x 3)
- Pooling layer (2 x 2)
- 512 weight candidates for dynamic parameters
- Cross entropy loss

Baseline - LF



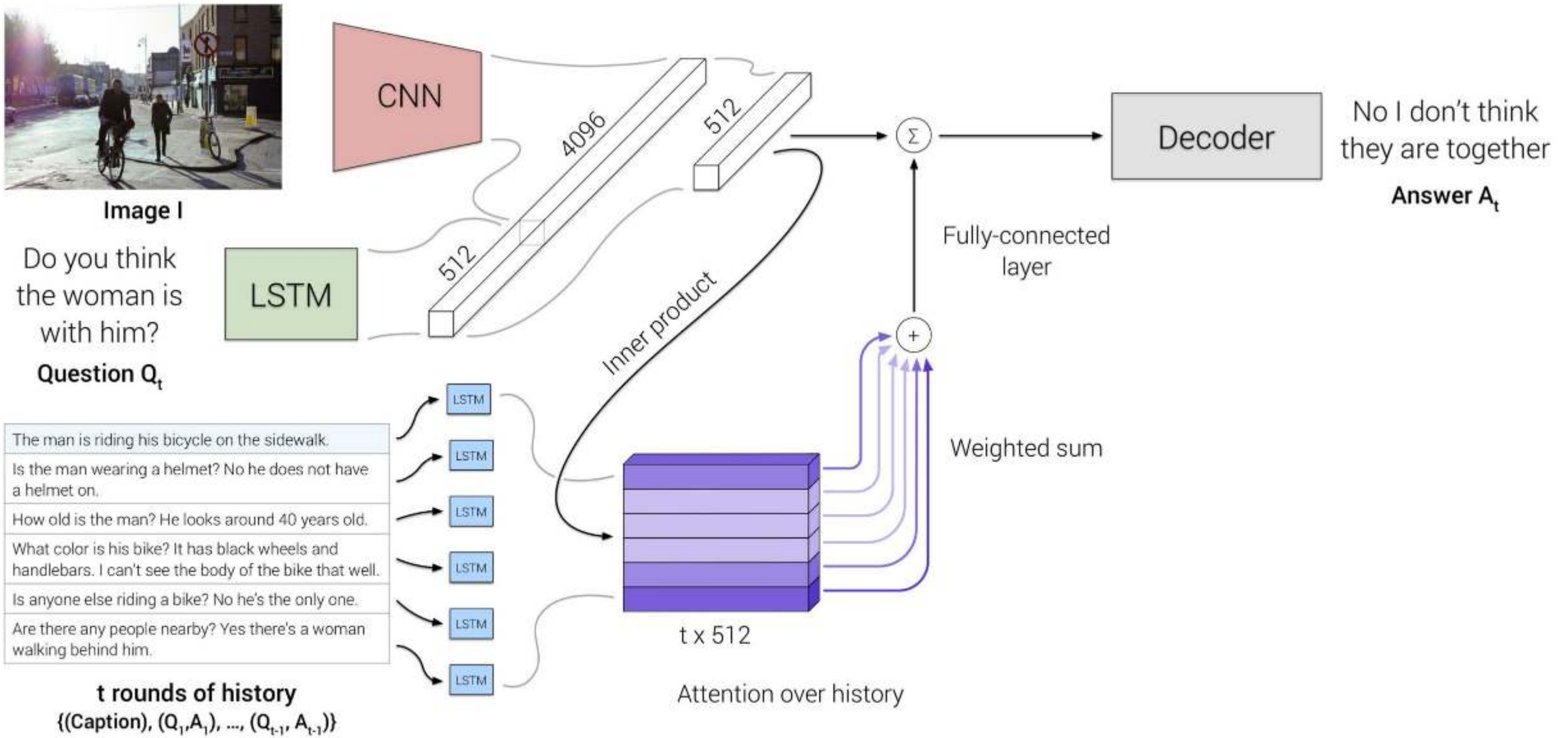
(a) Late Fusion Encoder

Baseline - HRE



(b) Hierarchical Recurrent Encoder

Baseline - MNE



(c) Memory Network Encoder

Results

Basemodel	+H	+SEQ	Accuracy
I	–	–	20.18
Q	–	–	36.58
	✓	–	37.58
LF [1]	✓	–	45.06
HRE [1]	✓	–	49.10
MN [1]	✓	–	48.51
ATT	–	–	62.62
	✓	–	79.72
AMEM	–	–	87.53
	✓	–	89.20
	–	✓	90.05
	✓	✓	**96.39**

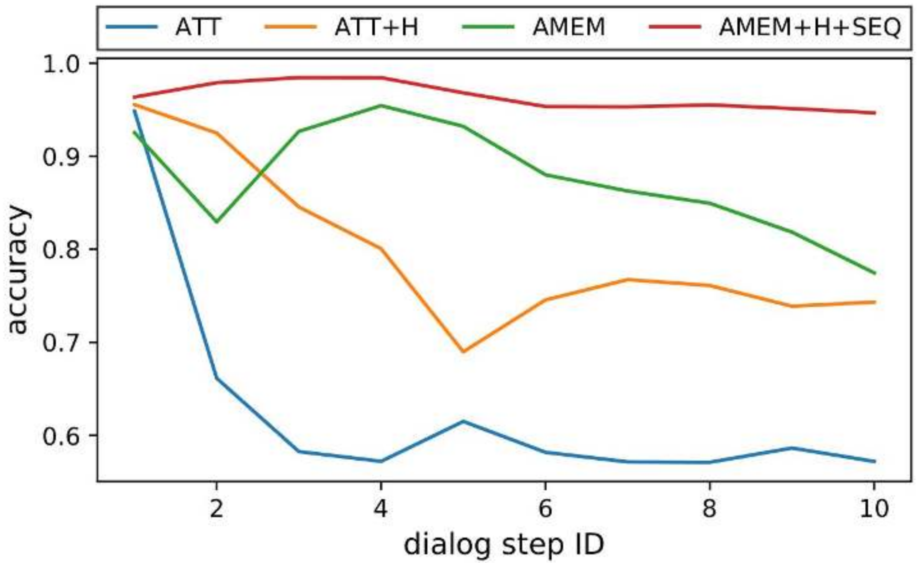


Figure 4: **Results on MNIST Dialog.** Answer prediction accuracy [%] of all models for all questions (left) and accuracy curves of four models at different dialog steps (right). +H and +SEQ represent the use of history embeddings in models and addressing with sequential preference, respectively.

Semantic interpretability

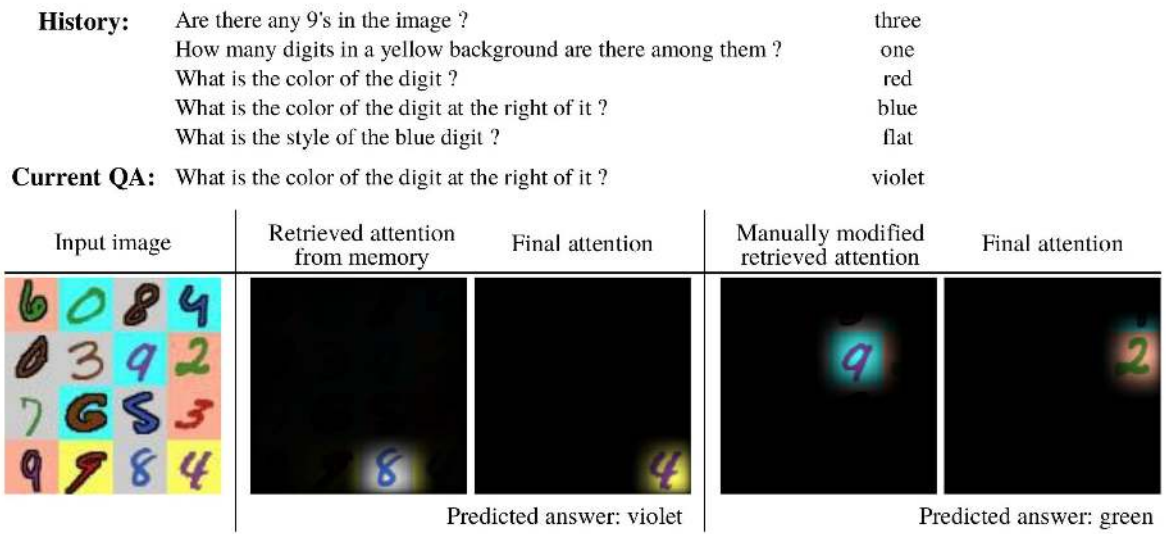
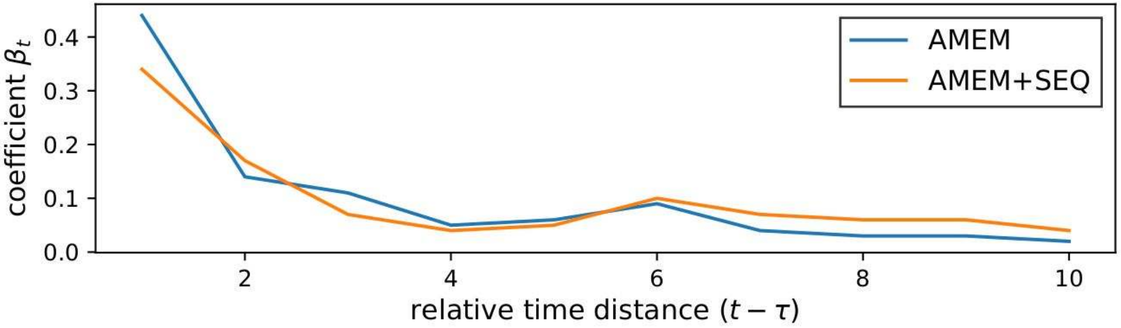


Figure 7: **Qualitative analysis on MNIST Dialog.** Given an input image and a series of questions with their visual grounding history, we present the memory retrieved and final attentions for the current question in the second and third columns, respectively. The proposed network correctly attends to target reference and predicts correct answer. The last two columns present the manually modified attention and the final attention obtained from the modified attention, respectively. Experiment shows consistency of transformation between attentions and semantic interpretability of our model.

Parameter analysis

- θ is consistently negative
  - Model prefers recent elements
- β_t also shows similar trend
  - Without bias W^{mem} puts too much focus on recent information

$$m'_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}}\boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau + \theta\,(t-\tau)$$



Parameter analysis

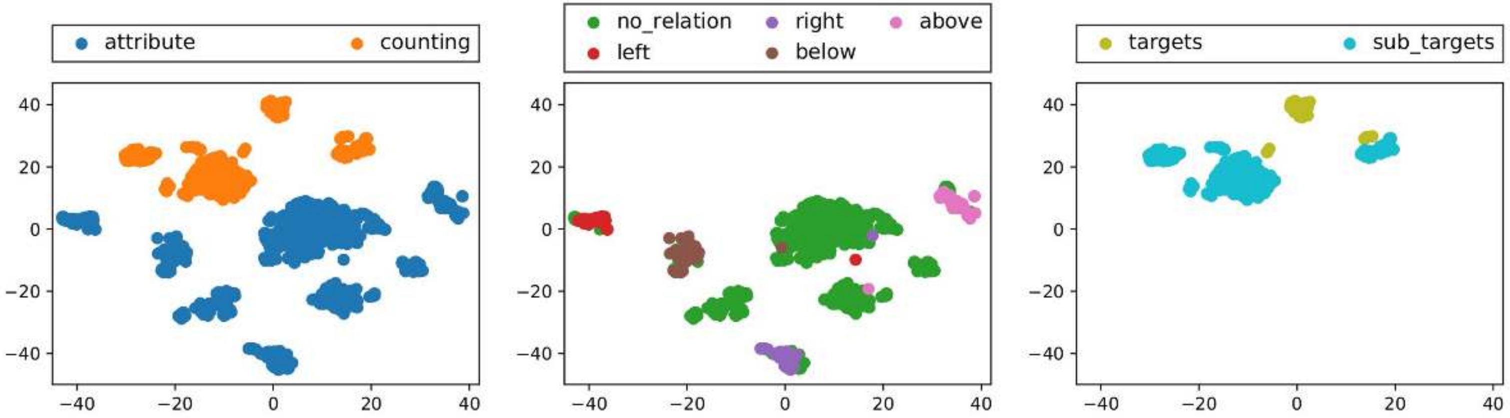


Figure 6: **Characteristics of dynamically predicted weights for attention combination.** Dynamic weights are computed from 1,500 random samples at dialog step 3 and plotted by t-SNE. Each figure presents clusters formed by different semantics of questions. (left) Clusters generated by different question types. (middle) Subclusters formed by types of spatial relationships in attribute questions. (right) Subclusters formed by ways of specifying targets in counting questions; cluster sub_targets contains questions whose current target digits are included in the targets of the previous question.

VisDial

- MS-COCO images + Caption + 10 Questions
  - Each question with 100 candidate answers
- Compared to MNIST Dialog
  - Answers are free form text
  - Contains fewer ambiguous expressions

Implementation details

- The initial history is constructed with the image caption
- Feature map extracted from conv5 layer of VGG-16 trained on ImageNet
- LSTM
  - Share weights for question and caption
  - Separate weight matrix for answers
- Word embedding (64 x 1) and Hidden state (128 x 1)
  - These dimensions are significantly lower than baseline models

Results

Table 1: **Experimental results on VisDial.** We show the number of parameters, mean reciprocal rank (MRR), recall@k and mean rank (MR). +H and ATT indicate use of history embeddings in prediction and attention mechanism, respectively.

Model	+H	ATT	# of params	MRR	R@1	R@5	R@10	MR
Answer prior [1]	–	–	n/a	0.3735	23.55	48.52	53.23	26.50
LF-Q [1]	–	–	8.3 M (3.6x)	0.5508	41.24	70.45	79.83	7.08
LF-QH [1]	✓	–	12.4 M (5.4x)	0.5578	41.75	71.45	80.94	6.74
LF-QI [1]	–	–	10.4 M (4.6x)	0.5759	43.33	74.27	83.68	5.87
LF-QIH [1]	✓	–	14.5 M (6.3x)	0.5807	43.82	74.68	84.07	5.78
HRE-QH [1]	✓	–	15.0 M (6.5x)	0.5695	42.70	73.25	82.97	6.11
HRE-QIH [1]	✓	–	16.8 M (7.3x)	0.5846	44.67	74.50	84.22	5.72
HREA-QIH [1]	✓	–	16.8 M (7.3x)	0.5868	44.82	74.81	84.36	5.66
MN-QH [1]	✓	–	12.4 M (5.4x)	0.5849	44.03	75.26	84.49	5.68
MN-QIH [1]	✓	–	14.7 M (6.4x)	0.5965	45.55	76.22	85.37	5.46
SAN-QI [10]	–	✓	n/a	0.5764	43.44	74.26	83.72	5.88
HieCoAtt-QI [15]	–	✓	n/a	0.5788	43.51	74.49	83.96	5.84
AMEM-QI	–	✓	**1.7 M (0.7x)**	0.6196	48.24	78.33	87.11	4.92
AMEM-QIH	✓	✓	2.3 M (1.0x)	0.6192	48.05	78.39	87.12	4.88
AMEM+SEQ-QI	–	✓	**1.7 M (0.7x)**	**0.6227**	**48.53**	**78.66**	**87.43**	**4.86**
AMEM+SEQ-QIH	✓	✓	2.3 M (1.0x)	0.6210	48.40	78.39	87.12	4.92

Strengths

- Novel attention mechanism which exploits visual attention history
- Achieves state-of-the-art performance with reduced number of parameters

Weaknesses

- MNIST Dialog dataset is not very representative of general visual dialog questions
- Should have evaluated performance over more ambiguous questions

Potential extensions

- Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning, Qi Wu et al, arXiv:1711.07613
    - Uses a combination of RL and GAN
    - Generates more human like answers
    - But does not use attention
- Use reinforcement learning approach
    - For example, Q-bot and A-bot
- Some combination of Attention, GAN and RL
- Use module networks for solving complex visual references

?