Pandabot: Multimodal Story Learning with Dynamic Memory Construction

Yu-Jung Heo, Eun-Sol Kim, Kyoung-Woon On and Byoung-Tak Zhang
Department of Computer Science and Engineering
Seoul National University
Seoul 08826, Korea
{yjheo, eskim, kwon, btzhang}@bi.snu.ac.kr

***Abstract*—We consider a challenging problem, video question and answering, based on multimodal sequential information. We suggest a memory-based machine learning algorithm which temporally combines three kinds of information, such as image stream, subscriptions and audio signals, of videos and infers an answer for a given question based on the constructed memory. Compared to conventional methods for question and answering, the suggested method can construct dynamic memory representations of multimodal sequential data. Also it can generate subjective answers rather than one-word objective answers for the questions of the human. For the experiments, 4000 pairs of question and answering data are collected by Amazon Mechanical Turk for five fairy tales of the child educational animation video. Experimental results on the collected question and answering dataset demonstrate meaningful improvements of the accuracy of the answers.**

Video question and answering is a challenging problem, because it is needed to infer high-level semantics based on the context of the video. In this paper, we propose a memory-based story learning method which can deal with multimodal sequential information about the story of video. Using the constructed multimodal dynamic memory, video question and answering problem is resolved.

The suggested method extends Dynamic Memory Network(DMN)[1] which focuses on language-based question and answering problems. DMN constructs episodic memory just using text information such as description and question sequence and generates relevant answers given a memory. Even though the model shows significant performance on question and answering, it is hard to apply to the video. Because, in video question and answering problem, not only text information explaining the situation but also visual and auditory information is crucial to understand overall context in given situation.

Therefore, in this paper, we combine three kinds of modalities, such as image stream, subscriptions, and audio signals, of the video to construct episodic memory. Also, as constructing each episode, we consider each question time stamp when a user asked each question watching the video. we focus on a point in time for question and answering, so that make the episodic memory dynamic and compacted.

A. Input Module

The input module combines three modalities, such as image stream, subscription and audio signals, of given video. Raw data for each modality is properly preprocessed to have high-level representations. We use a gated recurrent network(GRU)[2][3] to encode sequential information of preprocessed inputs I^t with time t. The equations of mechanism how to GRU network works are as follows.

$$z_t = \sigma(W^{(z)} I^t + U^{(z)} h_{t-1} + b^{(z)}) \quad (1)$$

$$r_t = \sigma(W^{(r)} I^t + U^{(r)} h_{t-1} + b^{(r)}) \quad (2)$$

$$\tilde{h_t} = tanh(W I^t + r_t \circ U h_{t-1} + b^{(h)}) \quad (3)$$

$$h_t = z_t \circ U h_{t-1} + (1 - z_t) \circ \tilde{h_t} = GRU(I_i^t, h_{t-1}) \quad (4)$$

B. Question Module

The question consists of a sequence of words. The question module encodes these words using the recurrent neural network same as input module. Hidden state of tth words of question is given by $q^{i_t} = GRU(embed[w_q^t], q^{i_{t-1}})$. Final hidden representation $q^{i_{T(q)}}$ is applied to construct episodic memory module m_i.

C. Episodic Memory Module

The episodic memory is constructed by iteration step which updates previous memory states over the input and question. It can be considered to apply change of attention on question and answering. The update equation of the episodic memory is $m_i = GRU(e^i, m_{i-1})$

D. Answer Module

The answer module generates hidden representations of answers $a = [a^1, a^2, ..., a^N]$, in which N is the number of answers. Each element of a means the final hidden representation of each answer, encoded by GRU module. tth hidden state is given by $a^{i_t} = GRU(embed[w_a^t], a^{i_{t-1}})$. With respect to final episodic memory module $m_{T(M)}$, the model output best answer $A^i = \min_i \|(m_{T(M)} - a^i)\|^2$.

E. Data

As experimental dataset, we used educational animations made by Pinkfong, No.1 grossing education app in 109 countries. We first selected five fairy tales, such as Hansel and Gretel, Snow White and the Seven Dwarves, The Little Mermaid, The Wolf and the Seven Sheep, and The Three Little Pigs. Then, we collected 800 question and answering pairs via

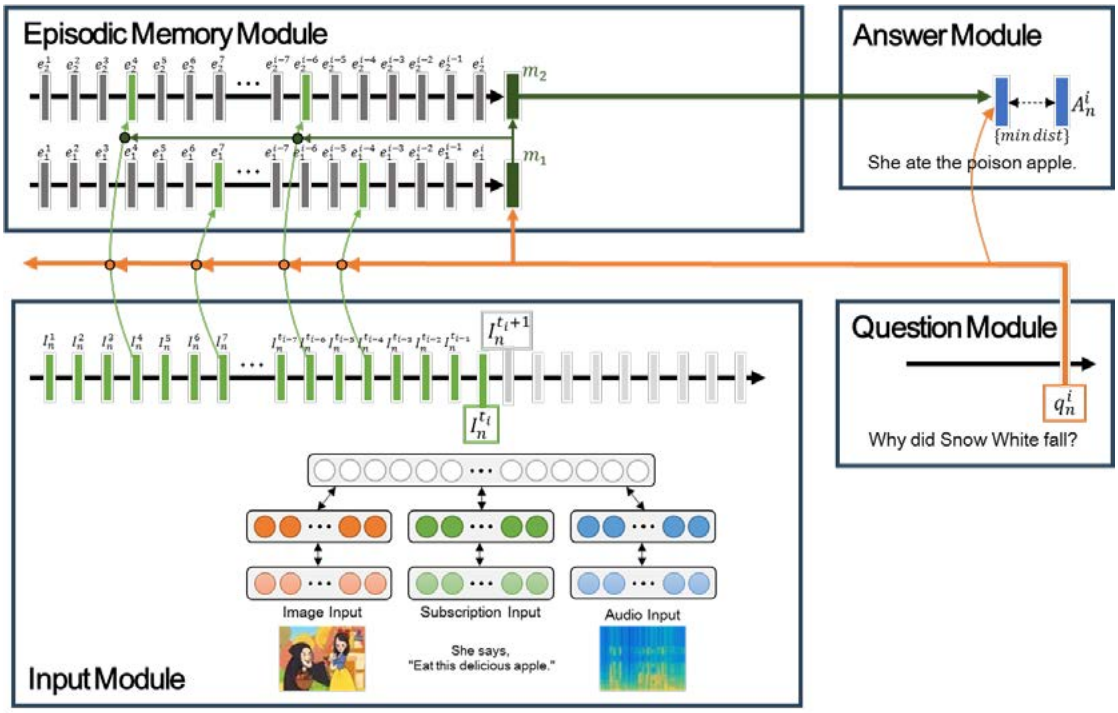

Fig. 1. An overall structure of Memory Network with dynamic memory construction for multimodal story learning

Amazon Mechanical Turk, a crowd-sourcing platform. Each QA pair has time stamp, which represents the timing of the contents related to the question. The question and answering pairs can be divided three types: first, question and answering about story(e.g., Why did Snow White fall? She ate the poison apple.), second, question and answering about visual information(e.g., What color is the poison apple? The apple is red.), third, question and answering about emotional feeling(e.g., What feeling does the Queen have? She is jealous.).

F. Preliminary results

We conducted preliminary experiments to verify suggested dynamic memory construction make episodic memory compacted. As the experiment setting, The text information of input and question module are embedded by pre-trained word embedding representation using GloVe[4]. We divided the collected dataset into two sets: one half for training set and the other for test set. Table I shows some examples of the test results.

G. Conclusion and Future works

In this paper, we proposed a memory-based story learning method for video question and answering problems. Also, we conducted preliminary experiments using dynamic memory construction and demonstrate the preliminary results. As the following experiments, we will show the results from multimodal sequential data of the video.

TABLE I
SOME EXAMPLES OF THE TEST RESULTS

TYPE	SENTENCE
Question	Does the hunter kill Snow White?
Answer	No he let her go into the Forest.
Prediction(Ours)	no hunter doesnt kill snow white.
Question	Why is the queen angry?
Answer	Because the mirror tells her Snow White is the most beautiful.
Prediction(Ours)	the mirror replied snow white is the most beautiful.

REFERENCES

[1] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," *arXiv preprint arXiv:1506.07285*, 2015.
[2] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–43.