Storybot: Story Learning from Cartoon Videos via Consecutive Event Embedding

Min-Oh Heo
Dept. of Computer Science and Engineering
Seoul National University
Seoul, Korea
moheo@bi.snu.ac.kr

Byoung-Tak Zhang
Dept. of Computer Science and Engineering
Seoul National University
Seoul, Korea
btzhang@bi.snu.ac.kr

***Abstract*— Household robots will live with humans together, and then they should know general temporal knowledge of everyday lives in various time scales. For learning the temporal knowledge of family members, observation-interaction-oriented dataset is fundamental, but, such data to understand contextual stories are not available yet. As an alternative, one of available data for learning by showing to robots is series of cartoon videos for young kids. This type of data has some advantages: 1) omnibus style: simple and explicit storyline in short, 2) narrative order use fabula (chronological sequencing), 3) limited number of characters and spatial environment. Here, we introduce the framework to learn stories of cartoon videos. To represent stories, we define an event as the concatenation of a continuous vector from a scene description sentence and one of a dialogue sentence. So, we collected description sentences for visual scenes by persons, and try to embed event vector onto latent space with favor of consecutive events using ranking loss. Using a series of approximately 200 episodes of cartoon videos named 'Pororo the Little Penguin', we visualize trajectory-like embedded space. We expect this approach to achieve the following goals: 1) easy to interpret the episodic context, 2) easy to approximate multi-scale missing events, 3) easy to infer blank scenes from the videos.**

Keywords—scene embedding; story learning; surrogate life data

I. INTRODUCTION

Recently, released are socially interactive household robots such as NAO, Pepper, and Jibo. In a few years, the robots will live humans together, and then they should know general knowledge of everyday lives including temporal knowledge in various time scales to understand human life patterns better. Since the knowledge of the family members is personalized and episodic, the robots should learn via observation and interaction in the environment. Ideal datasets for learning the temporal knowledge of family members are observation-interaction-oriented ones collected on real situated environments, but such data to understand contextual stories are not available in public yet. As an alternative, one of available data for learning by showing to robots is series of cartoon videos for young kids. They have some advantages: 1) omnibus style, which each episode has simple and explicit storyline in short, 2) narrative order mostly use fabula, which follows chronological sequencing of the events, whereas syuzhet is a term to designate the way a story is organized to enhance the effect of storytelling. 3) limited number of main characters and limited spatial environment. This is good for computational burden to need smaller complexity to learn. These properties are so desirable to provide the data similar to that of everyday lives in compact and explicit way.

In this paper, we assume the scenario as shown Figure 1: robots are watching cartoon videos on TV, and English subtitles are provided on the screen. That is, we pursue learning by showing to focus on story understanding.

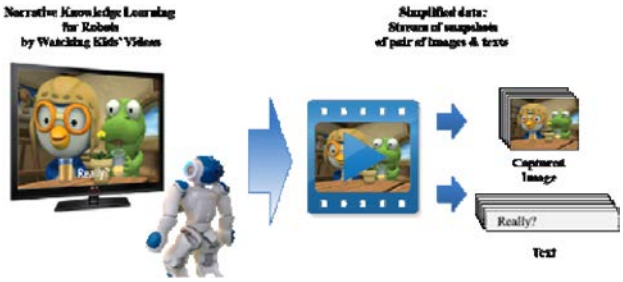

Figure 1. Scenario: Robot Learning by showing video series. As simplified data, a video stream converted to the stream of snapshots of pairs of images and texts.

We built new dataset from 3D animation videos for kids, entitled 'Pororo the Little Penguin', consisting of 16,066 scene-dialogue pairs created from the video of 20.5 hours in total length, 27,328 fine-grained descriptive sentences for scene descriptions.

II. METHOD

A. Scene Event definition

Story itself is so various to handle, so we define a story as sequences of events. In this framework, an event is shortest element to represent a story. As in Figure 1, we captured snapshots of images and texts pairs, so we assume each snapshot has one event. And, the information in the images can be exposed with description sentences.

As a result, to represent stories, we define an event as the concatenation of a continuous vector from a scene description sentence and one of a dialogue sentence. We use 'skip-thought vector[1]' to change continuous vectors from the sentences.

B. *Embedding consecutive events*

Recently, pair-wise ranking loss is used popular to learn association between multi-modal information to connect similar pairs. Mostly, hinge loss or triplet loss can be used. In this work, we used the following variant of triplet loss,

$$\min_g \sum_x \sum_k \max\{0, \alpha - s(x_t, x_{t+1}) + s(x_t, x_k)\}$$

x_t is one event vector at time index t, x_k is another event vector to be chosen randomly. This loss function drives event vector at t and t+1 close, arbitrary other vectors far.

III. EXPERIMENT AND VISUALIZATION

A. *Caption generation*

At first, to build generalized automated captioning module, we use neuraltalk2 built by Andrei Karpathy. It is composed of Convolutional neural networks and Long Short-term Memory(LSTM) to learn image and description sentence pairs. This tool also distribute pre-trained model with lots of data, but, if directly applied to our cases, it show bad results. So, we relearn our dataset on the models. The following figure shows some examples of the results.

Figure 2. caption generated with neuraltalk2

B. *Visualization*

To embed event vectors with favor of consecutive events, we use Multi-layer perceptron (MLP) as a mapping function from event vectors to latent space. And we used the loss function in 2. B, and visualize the latent space changing the number of nodes in each layer of MLPs.

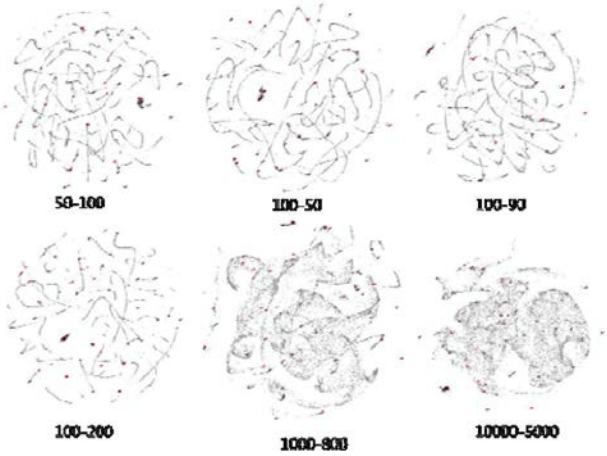

Figure 3. embedded event vector visualization with t-sne

IV. FINAL GOALS

We expect this approach to achieve the following goals: 1) easy to interpret the episodic context, 2) easy to approximate multi-scale missing events, 3) easy to infer blank scenes from the videos.

REFERENCES

[1] R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," In Advances in neural information processing systems (NIPS), pp. 3294-3302, 2015.

[2] L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research 9(Nov):2579-2605, 2008.