10-2018

Knowledge-aware multimodal fashion chatbot

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

You ZHOU

Yunshan MA

Richang HONG

Tat-Seng CHUA

Knowledge-aware Multimodal Fashion Chatbot

Lizi Liao[1], You Zhou[1], Yunshan Ma[1], Richang Hong[2],Tat-Seng Chua[1]
[1]National University of Singapore, [2]Hefei University of Technology
{liaolizi.llz,yuri.youzhou,mysbupt,hongrc.hfut}@gmail.com, chuats@comp.nus.edu.sg

ABSTRACT

Multimodal fashion chatbot provides a natural and informative way to fulfill customers' fashion needs. However, making it 'smart' in generating substantive responses remains a challenging problem. In this paper, we present a multimodal domain knowledge enriched fashion chatbot. It forms a taxonomy-based learning module to capture the fine-grained semantics in images and leverages an end-to-end neural conversational model to generate responses based on the conversation history, visual semantics, and domain knowledge. To avoid inconsistent dialogues, deep reinforcement learning method is used to further optimize the model.

ACM Reference Format:

Lizi Liao, You Zhou, Yunshan Ma, Richang Hong,Tat-Seng Chua. 2018. Knowledge-aware Multimodal Fashion Chatbot. In 2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3240508.3241399

1 INTRODUCTION

By offering an interactive and natural way for information seeking, multimodal chatbots are attracting increasing attention. However, most existing chatbot systems are limited in textual modality and largely overlook the inherent domain knowledge, thus result in unsatisfactory responses. In fashion domain, visual traits of products and knowledge such as matching style tips of apparel are essential for the system to generate substantive answers.

In this paper, we aim to develop an intelligent chatbot system which can generate responses in both textual and visual modalities, and enrich with fashion domain knowledge. An illustration of the system is shown in Figure 1. First, the system correctly understands the visual semantics inside the product image, and can thus properly respond to user's request about similar dresses in blue color, and manages to make accurate attribute modifications (e.g., changing from red color to blue) in the query. Second, the system has the capability to leverage domain knowledge such as fashion style tips to answer the user's question about whether the blue skater dress matches with the silver stilettos.

2 FRAMEWORK

We show the architecture of our knowledge enriched multimodal fashion chatbot in Figure 2. The first layer takes in user utterances in different modalities; the second layer deals with the intention identification layer which determines the routing of addressing each user question. In addition to classifying user intentions, it also determines which modality response should be generated. The third layer depicts different components for further question processing and response generation.

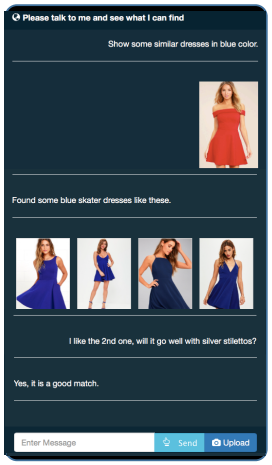

Figure 1: An example of a knowledge-aware multimodal fashion chatbot. The bot manages to understand the semantics of product image and modify the attributes during backend retrieval, offer matching suggestions for the user, and generate responses with different modalities.

Generally speaking, the system accepts multimodal utterances and then classifies user intentions to our predefined intention classes. The form of response (whether textual, visual or both) is also decided. For example, if the system detects that the user intention is to find products and a general category is present, the taxonomy traversal component will be activated. A set of subcategories will be returned to help the user narrowing down the search space. If it determines that the intention is about attribute and a specific product is detected, the slot filling component will be activated to generate the appropriate responses. There also exist some more complicated intentions and scenarios, for which the system resorts to the knowledge engine model to generate the responses.

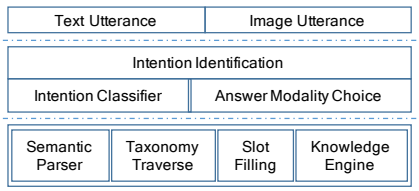

Figure 2: The overall architecture of the fashion chatbot.

Figure 3 illustrates the knowledge engine model. There are three major components. (1) In each turn, given a multimodal utterance, the model understands the semantics (category and attributes) of the product image via a taxonomy-based learning model. (2) Besides

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10.
https://doi.org/10.1145/3240508.3241399

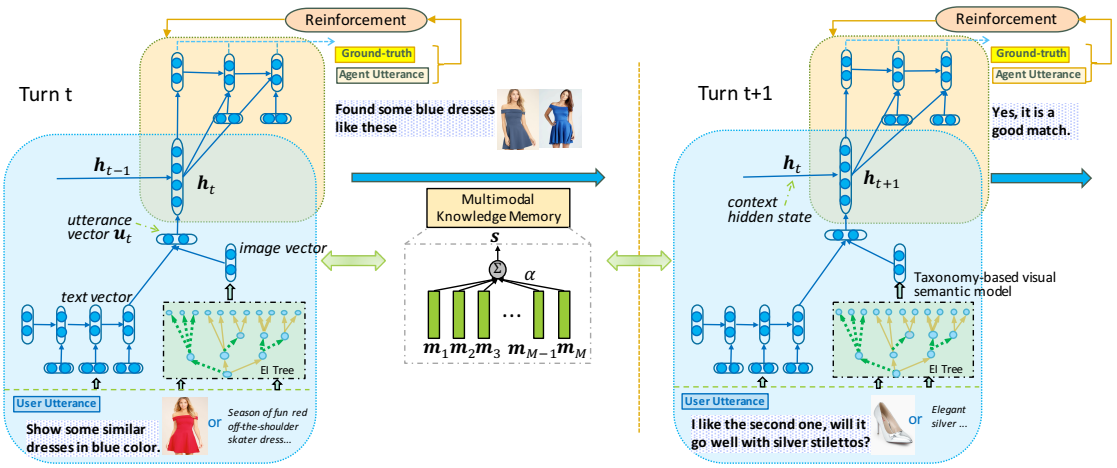

Figure 3: The knowledge engine model. Two turns of conversations are illustrated. The model leverages taxonomy-based visual semantic model to understand user utterances in different forms. It generates various forms of responses enriched with extracted domain knowledge. Deep reinforcement learning measures the goodness of a response through a reinforcement signal and optimizes the long-term rewards that characterize a good conversation.

modeling the utterances using the hierarchical recurrent encoder-decoder (HRED) network [4], the model employs an attention mechanism over the extracted style tips (e.g. blue jeans match with white shirts) and decides which knowledge is relevant to the context. The model then generates responses based on the conversation history and relevant knowledge stored in the memory network. (3) Based on the extended HRED backbone network, the system applies deep reinforcement learning that accounts for future rewards to optimize the neural conversational model using policy gradient methods.

3 IMPLEMENTATION

3.1 Intention Identification

User intentions in our model are classified into 10 categories: greeting, showing similar item, showing orientation, asking attribute, changing attribute, suited-for etc. We train our LSTM model for intention classification using the dataset published in [1], which consists of over 150K conversation sessions between shoppers and sales agents. We extend their intention categories to match with our scenarios. Regarding to the intention classification performance, our LSTM model manages to achieve a precision score of 91.23%.

3.2 Semantic Parser

We perform three steps in our semantic understanding: taxonomy positioning, image feature extraction, attribute detection. We use a dictionary to detect whether a utterance contains a category concept in our taxonomy. Such taxonomy location can be used for item recommendation and taxonomy traversal. For image input, we use taxonomy-based fashion concept learning model to understand the semantic information of the input images, and further recommend similar products to users. For more details, please refer to [2].

3.3 Knowledge Components Training

We train the knowledge engine model in two stages on the dataset published in [1]. In the first stage, we built on prior work of predicting a generated target utterance given the dialogue history using the knowledge enriched multimodal HRED model in a supervised fashion. Each response turn in the dataset was treated as a target and the concatenation of five previous utterances were treated as context. In the second stage, following the popular strategy in deep reinforcement learning training, we initialized the policy model using the extended HRED model trained during the first stage. This ensures that we start off with a much better policy than random and focus on a relatively good part of the search space. For more details, please refer to [3].

4 CONCLUSIONS

In this work, we present the knowledge enriched multimodal fashion chatbot that is specifically designed to help users in searching for products and matching styles. We demonstrate the system, present the techniques used, and share our experience in dealing with the challenges in this field. For future work, several points will be further explored to improve our chatbot including personalizing the conversation bot, transferring knowledge from other domains with rich data, and combining the task-oriented methods with non-task-oriented chatbots.

ACKNOWLEDGMENT

This research is part of NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative. This work is also supported in part by the project from the National Science Foundation of China under grant 61722204 and 61732007.

REFERENCES

[1] Amrita aha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. In *AAAI*. 696–704.
[2] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-seng Chua. 2018. Interpretable Multimodal Fashion Retrieval for Fashion Products. In *MM*.
[3] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *MM*.
[4] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical Neural Network Generative Models for Movie Dialogues. *arXiv preprint arXiv:1507.04808* (2015).