Building Goal-Oriented Dialogue Systems with Situated Visual Context

Sanchit Agarwal[1], **Jan Jezabek**[2*], **Arijit Biswas**[1], **Emre Barut**[1], **Bill Gao**[1], **Tagyoung Chung**[1]

[1] Amazon Alexa AI
[2] Hedgefrog Software
agsanchi@amazon.com, jezabek@gmail.com, {barijit,ebarut,shuyag,tagyoung}@amazon.com

Abstract

Goal-oriented dialogue agents can comfortably utilize the conversational context and understand its users' goals. However, in visually driven user experiences, these conversational agents are also required to make sense of the screen context in order to provide a proper interactive experience. In this paper, we propose a novel multimodal conversational framework where the dialogue agent's next action and their arguments are derived jointly conditioned both on the conversational and the visual context. We demonstrate the proposed approach via a prototypical furniture shopping experience for a multimodal virtual assistant.

Introduction

Goal-oriented dialogue systems enable users to complete specific goals such as booking a flight. The user informs their intent and the agent will ask for the slot values such as time and number of people before booking. Such traditional goal-oriented systems are often aware of the conversational context (Acharya et al. 2021; Wen et al. 2017; Liu et al. 2017; Shah et al. 2018), where users can refer to previous entities in the dialogue. However, there are numerous use cases where inference from a shared visual context (e.g. a device screen) is vital in fulfilling the users' goal. For instance, if a user browsing to shop for chairs on a speech enabled smart TV says *"What is the price of the black checkered one?"*, AI needs to identify the right product based on its visual characteristics and then respond with the price. Moreover, in certain cases one might need to infer from past visual context, e.g. when a user wants to compare items from the current screen with those shown on earlier screens by asking questions such as *"Is this cheaper than the green t-shirt you showed earlier?"*. Currently, majority of the commercially available conversational AI systems do not fully account for such visual context, and they are unable to fulfill such goals.

In this demo, we show a novel multimodal goal-oriented dialogue system that can reason over the current and historical screen context, as well as the conversational context, in order to complete users' goal. Users can refer to on-screen visual entities by attributes such as color (e.g., "the white one") or shape (e.g., "the one shaped like an airplane"). They can also refer to the visual elements using associated metadata such as rating (e.g., "five-star one") or by their relative positions (e.g., "middle one"). The dialogue agent can automatically resolve the relevant visual entities from the context and determine the user's intent. For example, *"Zoom in on the red-striped shirt"* calls the *ZOOM* action with the argument being the entity representing the "red-striped shirt" in the visual context. We enable multimodal understanding capabilities by introducing a new visual grounding model that can reason over the visual context given the user query and populate API arguments with visual entities.

[*]Work done while at Amazon Alexa AI.

Approach

Our proposed model for grounding the user query with respect to visual context consists of three main components: a query encoder, a visual entity encoder, and a candidate scorer. At a high level, we (i) first encode the query and each visual entity in the candidate set, which consists of both the entities currently on the screen and the entities seen in the previous interactions; (ii) compute a score for each (query, visual entity) pair; and (iii) finally choose the highest scoring entity.

Model Architecture

We present the full architecture in Figure 1. We describe its various components in the following subsections.

Query Encoder The query consists of the dialogue context and the current user utterance. Both the dialogue context and the current user utterance are encoded with (different) bi-directional LSTMs. The final query representation is obtained by concatenating the final hidden states from the LSTMs.

Visual Entity Encoder The visual entities of the displayed products consist of three components: (i) the metadata, which contain information such as item/brand name and rating; (ii) the image of the product; and (iii) contextual attributes of the item, such as its location on the screen, current visibility, and when was the last time the object was seen/mentioned by the user. All of these are crucial for a seamless experience, as users might refer to products by one, or a combination, of the above features (e.g., *"Add the*

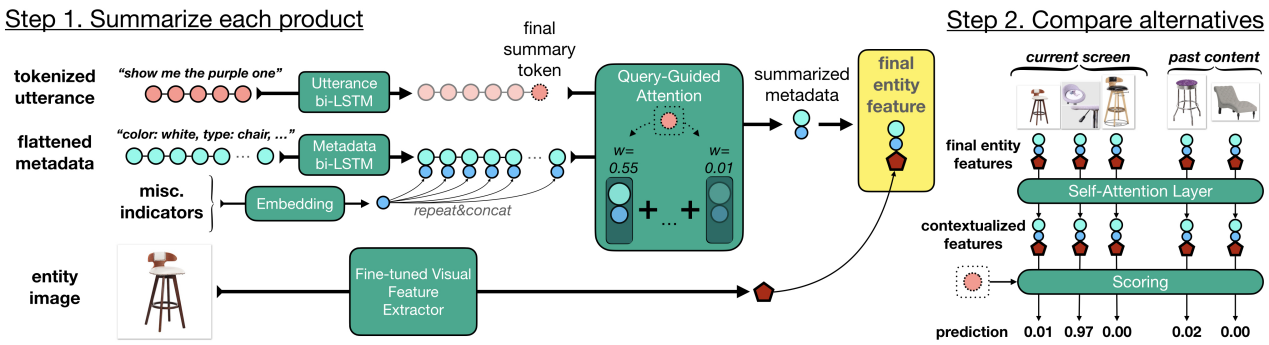

Figure 1: Overview of the proposed model architecture.

striped chair to my cart", and *"What was the price of the previous green chair?'*).

We encode the metadata using a bi-directional LSTM over a flattened sequence of metadata property names and their values. The flattened metadata sequence can be arbitrarily long, since there could be large number of attributes. To help the model focus on the desired attributes, we utilize query-guided attention, where the query embedding attends to each token in the metadata sequence, determines a weight for each, and combines them to produce a query-attended metadata encoding. The image is encoded via a pre-trained ResNet-50 model, where the last couple of layers are fine-tuned. For encoding the other contextual information, we utilize a combination of sinusoidal (Vaswani et al. 2017) and dense embeddings, for capturing spatial/temporal and visibility aspects, respectively. Embeddings from the three components are concatenated to form a joint representation.

Finally, in order to contextualize each visual entity with respect to the other entities, we add a self-attention layer on top of the joint representation to produce the final representation. This allows the model to perform comparative reasoning (e.g. *"Is the most expensive one available in green?"*).

Candidate Scorer The final scorer is a bi-linear attention layer that computes an attention score between the query and each visual entity in the candidate set. The attention scores are used as relevance scores, and the entity with the highest score is returned as the output. The full model is trained by minimizing the cross-entropy of the chosen label and the prediction scores.

Simulator

In order to increase the generalization capabilities of our model by complementing our training data, we build a multimodal dialog simulator that generates synthetic conversations. This synthetic data increases the variation of visual context in the training data and provides examples of interactions spanning multiple turns. The simulator randomly generates screen layouts and simulates user interactions with entities on the screen or with other entities that have been mentioned previously in the dialog. Simulated interactions contain questions about properties of products (*"what is the material for the left one"*), comparison between products, as well as utterances for taking certain actions on them (*"add the red one to the cart"*).

Model	Accuracy (%)	
	Simulated Test Set	MTurk Test Set
Proposed Model (Only current screen context)	84.89	84.23
Proposed Model (Screen context from last 3 turns)	97.73	85.13

Table 1: Performance of the proposed model with and without the past visual context.

Experiments and Discussion

The product catalog for our dataset is built using a small subset of Amazon catalogue with close to 50,000 furnitures. We generate the simulated dataset using the multimodal dialogue simulator as described above. In order to collect more realistic and diverse data, we augment the simulated data with an MTurk collection where we display various items to the MTurkers and ask them to refer to one of the items on the screen using a natural language query. We collect 95,000 utterances, which we split into $8:1:1$ for training, development, and test without overlap.

Table 1 shows the performance of the proposed visual grounding model on simulated and MTurk test sets. We train two variants of the proposed model; in the first, the candidate set is derived only from the current screen, and in the second, the candidate set also includes items from previous screens. MTurk collection utterances only contain references to the current screen, therefore the performance on MTurk test set is largely unchanged whether or not we train the proposed model with historical visual context. However, on the simulated set, including historical entities gives significant improvement, primarily in cases that reference a previously shown entity, e.g., *"Is the blue one cheaper than the previous red one"*. Overall, our proposed model achieves an accuracy of 85% on a realistic and relatively difficult MTurk dataset.

References

Acharya, A.; Adhikari, S.; Agarwal, S.; Auvray, V.; Belgamwar, N.; Biswas, A.; Chandra, S.; Chung, T.; Fazel-Zarandi, M.; Gabriel, R.; Gao, S.; Goel, R.; Hakkani-Tur, D.; Jezabek, J.; Jha, A.; Kao, J.-Y.; Krishnan, P.; Ku, P.; Goyal, A.; Lin, C.-W.; Liu, Q.; Mandal, A.; Metallinou, A.; Naik, V.; Pan, Y.; Paul, S.; Perera, V.; Sethi, A.; Shen, M.; Strom, N.; and Wang, E. 2021. Alexa Conversations: An Extensible Data-driven Approach for Building Task-oriented Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, 125–132. Online: Association for Computational Linguistics.

Liu, B.; Tur, G.; Hakkani-Tur, D.; Shah, P.; and Heck, L. 2017. End-to-End Optimization of Task-Oriented Dialogue Model with Deep Reinforcement Learning. *NIPS 2017 Workshop on Conversational AI*.

Shah, P.; Hakkani-Tür, D.; Tür, G.; Rastogi, A.; Bapna, A.; Nayak, N.; and Heck, L. 2018. Building a Conversational Agent Overnight with Dialogue Self-Play. *arXiv preprint arXiv:1801.04871*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wen, T. H.; Vandyke, D.; Mrkšíc, N.; Gašíc, M.; Rojas-Barahona, L. M.; Su, P. H.; Ultes, S.; and Young, S. 2017. A network-based end-to-end trainable task-oriented dialogue system. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1: 438–449.