

Towards Building Large Scale Multimodal Domain-Aware Conversation Systems

Amrita Saha
IBM Research AI and
Indian Institute of Technology
Madras, India
amrsaha4@in.ibm.com

Mitesh M. Khapra
Indian Institute of Technology
Madras, India
miteshk@cse.iitm.ac.in

Karthik Sankaranarayanan
IBM Research AI
kartsank@in.ibm.com

Abstract

While multimodal conversation agents are gaining importance in several domains such as retail, travel etc., deep learning research in this area has been limited primarily due to the lack of availability of large-scale, open chatlogs. To overcome this bottleneck, in this paper we introduce the task of multimodal, domain-aware conversations, and propose the MMD benchmark dataset. This dataset was gathered by working in close coordination with large number of domain experts in the retail domain. These experts suggested various conversations flows and dialog states which are typically seen in multimodal conversations in the fashion domain. Keeping these flows and states in mind, we created a dataset consisting of over 150K conversation sessions between shoppers and sales agents, with the help of in-house annotators using a semi-automated manually intense iterative process. With this dataset, we propose 5 new sub-tasks for multimodal conversations along with their evaluation methodology. We also propose two multimodal neural models in the encode-attend-decode paradigm and demonstrate their performance on two of the sub-tasks, namely text response generation and best image response selection. These experiments serve to establish baseline performance and open new research directions for each of these sub-tasks. Further, for each of the sub-tasks, we present a ‘per-state evaluation’ of 9 most significant dialog states, which would enable more focused research into understanding the challenges and complexities involved in each of these states.

Introduction

The recent progress with deep learning techniques for problems at the intersection of NLP and Computer Vision such as image captioning (Xu et al. 2015), video description (Yu et al. 2016), image question answering (Antol et al. 2015), video question answering (Zeng et al. 2016; Maharaj et al. 2016), is owed largely due to the availability of large-scale open datasets for their respective tasks. However, even though there is a growing demand for chatbots that can converse using multiple modalities with humans in several domains such as retail, travel, entertainment, etc. the primary hindrance for deep learning research in this area has been the lack of large-scale open datasets.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Though there has been recent work (Serban et al. 2016b; Yao et al. 2016; Serban et al. 2016a) with different conversation datasets (Lowe et al. 2015; Vinyals and Le 2015; Ritter, Cherry, and Dolan 2010), the mode of interaction there is limited to text conversations only. While multimodal, human-to-human conversation transcripts (e.g. between shoppers and salespersons) might be available in industry settings, they are both limited in scale and proprietary, thus hindering open research.

In this paper, we attempt to alleviate these challenges by developing a large-scale multimodal conversational dataset in the retail domain that embodies the required generic capabilities for such autonomous agents. Since the actual transcripts are both limited and proprietary, we conducted a series of interviews with a large number of retail salespersons from the fashion domain and developed the dataset interactively with them in the following semi-automatic manner. The domain experts described in detail, various phases of the sales process which were materialized into 84 states for a conversational agent. Each such state had a specific intent (e.g. a shopper wanting to see more items similar to a specific one identified, or asking for a style tip or about the latest trends being endorsed by celebrities, etc). Corresponding to each such intent, a collection of representative utterance patterns involving both text and images were designed. Each of these states exhibited multimodality (i.e. involving text and images) in both the utterance/response as well as the dialog context. The domain experts described various possible sales flows where the customer went from one state to another, which were captured by transitions of an expert model automata between these states. The experts then inspected the outputs of controlled runs of the automata, provided feedback, which was thereafter employed to further refine the automata. Proceeding in this *manually-intensive* and *iterative* manner under the supervision of domain experts, we produced a large-scale dataset of over 150K multimodal dialogs capturing a wide variety of conversational sessions exhibiting the challenges described earlier. Note that such a data collection could not have been achieved with typical crowdsourcing platforms such as Amazon Mechanical Turk using ordinary crowd workers since it was necessary to be performed under the supervision of fashion sales experts exhibiting domain-

SHOPPER: Hello

AGENT: Hello, is there anything i can help you with today?

SHOPPER: show me some espadrilles in beige/blue within \$200.

AGENT: Sure. let me just quickly browse through my catalogue



AGENT:

AGENT: What you can see in the fifth item, michael michael kors brings you an effortless weekend look with the wrapped espadrille heel and crisscross styling of these dana flats. And what you see in the 3rd item, steve madden, the company, was founded by steve madden, the man, in 1990 in brooklyn, where he started out making trend-advancing shoes one pair at a time.

SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and the 5th



AGENT:

AGENT: The 2nd one, viscata espadrilles are 100% authentic and handcrafted in spain inspired by the mediterranean, viscata shoes are lightweight and made of natural jute rope, premium leathers, suedes and woven cloth.

SHOPPER: Can you list images of the 1st result from all orientations?



AGENT: Image from the front,right and back orientations respectively

SHOPPER: And the 3rd one?



AGENT: Image from the front,right,back and left orientations respectively

SHOPPER: I like the 4th image. Show me something like it but in style as in this image,



within my budget.



AGENT:

AGENT: Regarding the first item, platform espadrille slip on sam edelman shoes epitomize chic comfort.

SHOPPER: I am keen on seeing something similar to the 1st image but in a different sole material



AGENT: The similar looking ones are

SHOPPER: See the 1st espadrilles. I wish to see more like it but in silver coloured type



AGENT:

AGENT: In the third one, cobblerz presents these black coloured casual shoes, which will catch your fancy at once. And about the fifth item, be the cynosure of all eyes with this pair of silver coloured sandals by next.

SHOPPER: Will these espadrilles suit office style?

AGENT: Yes

SHOPPER: Are the products in the 2nd and 3rd images suited for softner washing?

AGENT: No

SHOPPER: And for hand wash?

AGENT: Yes

SHOPPER: Show me more images of the 3rd product in some different directions



AGENT: Image from the front,right,back and left view points respectively

SHOPPER: What about the product in the 5th image?



AGENT: Image from the front and back orientations

SHOPPER: Does celebrity cel_34 usually wear the kind of espadrilles in the 3rd image?

AGENT: No. But Celebrities cel_24 and cel_116 endorse this type of espadrilles

SHOPPER: What other apparels and accessories are usually endorsed by celebrity cel_116?

AGENT: loose fit, snake skin upper material, animal-texture pattern clogs is often endorsed

SHOPPER: Will the 5th result go well with a large size messenger bag?

AGENT: No

SHOPPER: What is the 1st images upper material ?

AGENT: The espadrilles in the 1st image has synthetic upper material

SHOPPER: Thank you for showing me all this, i guess i will leave

AGENT: Sure you are welcome

Figure 1: Example Dialog session between a shopper and an agent

specific understanding.

One might argue that such *semi-simulated* conversations do not capture the flavor of *natural* conversations. However, this is not entirely true for the conversations in our dataset. First, the constant supervision and inputs from domain experts ensures that the conversation remains grounded and is close to real world conversations in the fashion domain. Secondly, the actual utterances in the dataset are not machine generated but solicited from in-house annotators. In particular, for any given state (say, “express-interest-in-apparel”), in-house annotators were asked to provide natural language sentences corresponding to the state (for example, “I think the second shirt looks great”). This is in contrast to the recent dialog datasets released by (Bordes and Weston 2016) where the utterances are also machine generated. Also note that while there are 84 states in the conversation (which is a reasonable number given that the conversation is restricted to a specific domain), the fashion experts suggested various ways in which these states combine in natural conversations which results

in a large number of paths that the conversation can follow. Again, since these paths were supervised by fashion experts, they are very similar to natural conversations in the domain. Collecting completely *natural* real world conversation data at this scale is clearly infeasible and a data collection strategy of close-human machine interaction is perhaps a reasonable middle ground.

An example of a typical multimodal conversation in our dataset involving both text and images, and exhibiting domain knowledge is shown in Figure 1. Notice that the *response* generated by the system (or agent) can either be text (for example, see the 3rd turn inside the highlighted portion in red) or a set of images (as seen in the 1st turn there). Similarly, at every point, the context available to the system is multimodal which the system has to process to generate a coherent response. For example, to give a relevant response in Turn 2 inside the box the agent needs to first pay attention to the words “espadrille”, “silver” and the sentiment which are mentioned in the text and then understand the style (“strapped ankle”) from the image to fetch images for “silver espadrille

with *strapped ankle*” in response to the given context.

The body of work most relevant to ours is Visual QA (Antol et al. 2015) involving a single question and response, the work of (Mostafazadeh et al. 2017; Das et al. 2016) involving a sequence of QA pairs with a single image forming a dialog, and the work of (de Vries et al. 2016) which focuses on scene understanding and reasoning from a single image. There are a few key differences between these datasets and our work as demonstrated by the example above. First, in these cases, all questions in a sequence pertain only to a single image. Secondly, their responses are always textual. However, as is the case with natural conversations amongst humans, in our work, (i) there could be multiple images providing context, (ii) these context images could change across turns during the course of the conversation, and (iii) the response at each step can be text or image(s) or both. We would also like to mention that there is some recent work by (Yu-I Ha 2017) wherein they use fashion data from Instagram to analyze visual features of fashion images and correlate them with likes and comments on social media. Unlike our work they do not focus on multimodal conversations.

Finally, in this paper, we propose a baseline framework to model the agent’s responses in such multimodal conversations. In particular, we propose hierarchical dialog models for the textual and image response as two separate tasks and empirically estimate the feasibility of these tasks. We also discuss limitations that open new directions for research into these and multiple other tasks enabled by this new dataset. The following summarizes the main contributions of this work:

- We introduce a Multimodal Conversation task which is significantly distinct from the sequential Visual QA driven dialog tasks mentioned before
- We introduce a large dataset for this task and define several research tasks for evaluating it
- We propose baseline multimodal encoder decoder models for two such tasks and define appropriate metrics for evaluating these tasks

Multimodal Dialogs (MMD) Dataset

As mentioned in the previous section, a key contribution of this paper is a large-scale dataset of 2-party dialogs that seamlessly employ multimodal data in their utterances and context and also demonstrate domain-specific knowledge in their series of interactions. Towards this goal, in this section, we first describe the methodology employed for collecting this dataset and then explain in detail the various sub-tasks exhibited by the dataset that open up new research problems.

Data Collection Methodology

The data collection done in close coordination with a team of 20 fashion experts, primarily consisted of two steps, (i) curation and representation of a large-scale domain knowledge, and (ii) developing a large collection of multimodal conversations, each consisting of a series of interactions employing this knowledge. We next proceed to describe these two steps in detail.

Domain Knowledge Curation Through our series of interviews with the domain experts, we observed that a lot of the complexity in a natural conversation in this domain comes from the background knowledge that both the expert agent and the shopper employ in their conversation. The expert’s domain knowledge is multitude in nature, varying from knowledge about which attire goes well with which accessory, to which celebrity is presently endorsing which kind of fashion items, or what kind of look is better suited for which occasion. Therefore, the first step in our data collection process was to curate this domain knowledge from unstructured multimodal content on the web at scale and represent them in a machine consumable manner. This process involved a series of steps as enumerated below:

1. Crawling over 1 Million fashion items from the web along with their available semi/un-structured information and associated image(s).
2. Parsing different types of domain knowledge from the free-text information, and curating them in a structured form after a round of manual inspection by domain experts
 - 2a. Creating a hand-crafted taxonomy of the different types of fashion items. For example, *man > apparel > layer-2-lower-body > trouser > formal-trousers, dressed pants* i.e. *formal-trousers* is synonymous to *dressed pants* and is a type of *trouser* which is again a type of *layer-2-lower body* apparel. Each taxonomy entry has a synonym-set (called “synset”). With the help of domain experts we collected 282 such fashion “synsets” for men and 434 for women
 - 2b. Identifying the set of fashion attributes relevant (especially for the purpose of shopping) to each of the fashion synsets. Overall 52 such attributes (like color, pattern, style, price, wash-care information) were identified by domain experts, where 45 of them are visual attributes and remaining are meta-data attributes about the synset (e.g. wash-care information, price, seller ranking)
 - 2c. Seeding the attribute lexicons with a set of realistic values provided by the domain experts
3. Parsing the semi-structured catalog into a single unified structure of the form $\langle \text{synset}, \{ \text{attribute}: \{ \text{attribute values} \} \rangle, \{ \}$ denoting a set
4. Constructing a distribution of attributes and values for each of these synsets, from the structured catalog data curated in step 3 and filtering them through a close manual inspection by the domain experts
5. From the unstructured product description in the catalog, spotting and extracting style-tip information (e.g. black trousers go well with white shirt)
6. Creating fashion profiles for celebrities based on the type of clothes and accessories worn or endorsed by them. Since the profile information for real celebrities was proprietary, we generated profiles of imaginary celebrities by simulating a distribution of fashion synsets that each of these celebrities endorse, and a further distribution of fashion attributes preferred by these celebrities for each of these synsets. Note that doing so does not affect the generality of the dataset technically. Statistics about the final domain knowledge curated using this semi-automated methodology are tabulated in Table 1

Knowledge Base Statistics		Examples
#Items Crawled	1.05M	-
Avg. #Images per item	4	-
#Fashion Synsets	716	shirt, trouser, tuxedo, loafer, stilettoes, sunglasses, handbag, hat
#Fashion Attributes	52	color, pattern, material, brand, style, sleeves, price, wash-care,
#Visual Fashion Attributes	45	color, pattern, material, style, neck, sleeves, length, sole type, closure
#Meta-Info Attributes	8	price, wash-care, product ranking, brand, size, occasions
Avg. #Fashion Attribute per Synset	16	-
Avg. #values per attribute	500	-
#Coarse-Grained StyleTips (Synset, Synset association)	8871	shirt & trouser, tshirt & sneakers, tuxedo & cufflinks, suit & formal shoes, kurta & jeans
#Fine-Grained StyleTips (Attribute Synset, Attribute Synset association)	350K	white shirt & black trousers, light t-shirt & dark jacket, black gown and silver stilettoes
#Celebrity profiles	411	Celeb1 likes vegan-leather shoes, monochrome pastel shaded t-shirts and polyester jackets
Avg. #Synsets endorsed by a celebrity	4	Celeb1 endorses lehenga, saree, blouse, kurti, sandals, stilettos.
Avg. #Celebrities endorsing a synset	15	-
#Synsets endorsed by celebrities	252	-

Table 1: Domain Specific Knowledge Base Statistics

Gathering multimodal dialogs During the interviews, the domain experts described in detail various phases of the sales process. For example, a dialog between the sales agent and a shopper who visits an e-commerce website with the objective of either buying or browsing one or more fashion items begins by the shopper providing their shopping requirements to the agent. The agent then browses the corpus and comes back with a multimodal response (i.e. with a set of images that satisfy the shopper’s constraints and/or some associated text). Now, using this response the shopper provides feedback or modifies their requirements. Through this iterative response and feedback loop the shopper continues to explore their items of interest, adding chosen items to their shopping cart. The session continues until they either choose to exit without a purchase or culminates with the shopper buying one or more items. Note that at various steps during such a conversation, the response of the current step of the dialog is based on inference drawn from an aggregate of images and text in the unstructured dialog context as well as a structured background domain knowledge (which is again multimodal in nature).

The domain experts described each of these various possible types of states involved in the conversations, a subset of which are shown in Table 2. This was mapped to an expert model automata which consisted of a total of 17 state types for the shopper covering 84 states. Each such state had a specific intent and corresponding to them, a collection of representative utterance patterns involving both text and images were designed by us along with the experts. Each such state would exhibit the following 3 features: (a) multimodality of utterance/response: shopper’s utterance and the agent’s response could involve either text or image or both, (b) multimodality of context: shopper’s utterance would employ the context of the conversation which would include both the text history and a number of images and (c) combination of the structured domain knowledge and the unstructured dialog context, both being multimodal.

Id	Automata State Type	State Description
1	greeting	Shopper greets
2	self-info	Shopper gives information about him/herself
3	give-criteria	describes his requirements
4	show-image	Agent shows relevant responses for the Shopper’s query
5	give-image-description	Agent generates short description of the product, using visual and catalog information
6	Like/Dislike specific items / overall items, show-more	Shopper expresses negative or positive preference specifically towards one or more items previously or currently shown, or a overall general preference towards all the items and optionally shows a new image to possibly modify his requirements and wants to see more
7	show-orientation	Shopper wants to see an item from different orientations
8	show-similar	Shopper wants to see similar to a particular item
9	goes-with	Shopper asks for style-tip
10	ask-attribute	Shopper asks about the attributes of the items shown
11	suited-for	Shopper asks about what are suited for that item
12	sort-results	Shopper wants to sort the result set by some attribute
13	filter-results	Shopper wants to filter the results based on some attribute
14	celebrity	Shopper asks questions relating to some celebrities and his fashion items of interest
15	switch-synset	Shopper wants to switch back to the type of fashion synset he had seen previously
16	buy	Shopper wants to buy one or more items
17	exit	Shopper wants to exit

Table 2: Details of example Automata State-Types as described by domain experts. 14 of these state-types in turn have multiple different states, thus yielding overall 84 automata states to be used in the dialog

The domain experts then provided a large number of possible sales process flows of customers proceeding from one state to another. These transitions were captured by the automata between these states with expert designed transition probabilities. The domain experts then inspected the outputs of small runs of the automata and provided feedback. This feedback was then incorporated to further refine the expert automata whose runs were again inspected by the experts. This iterative process was manually-intensive, and required close coordination with the domain experts. Following this process, we produced a large-scale dataset of over 150K multimodal dialogs.

Qualitative Survey To ensure that the dataset is representative and not biased by the specific fashion experts interviewed, we conducted a survey of the dataset by involving a different set of 16 fashion experts. They were asked to evaluate both whether the text portions of the dialog are natural sounding and meaningful and whether the images in it are appropriate. The survey was conducted with a randomly sampled set of 760 dialog sessions and the experts were asked to provide an overall rating between 1 to 5 (with 5 being most realistic).

Two types of errors were documented: (i) minor error being conversational mistakes (e.g. grammatical and phrasing

Dataset Statistics	Train	Valid	Test
#Dialogs(chat sessions)	105,439	22,595	22,595
Proportion in terms of dialogs	70%	15%	15%
Avg. #Utterances per dialog	40	40	40
#Utterances with shopper’s questions	2M	446K	445K
#Utterances with agent’s image response	904K	194K	193K
#Utterances with agent’s text response	1.54M	331K	330K
Avg. #Positive images in agent’s image response	4	4	4
Avg. #Negative images in agent’s image response	4	4	4
Avg. #Words in shopper’s Question	12	12	12
Avg. #Words in agent’s text Response	14	14	14
Avg. #Automata states per dialog	15	15	15
Avg. #Dialogs having a particular automata state	20,246	4,346	4,335
Avg. #Automata state-types per dialog	13	13	13
Avg. #dialogs having a particular state-type	59,638	12,806	12,764
#Automata states	84	84	84
#Automata state types	22	22	22
Vocabulary Size (threshold frequency >=4)	26,422	-	-

Table 3: Multimodal Dialog Dataset Statistics

error), (ii) severe error being logical mistakes (e.g. deductions errors in generating the image or text response, incorrect understanding of the shopper’s question, wrong fashion recommendation, etc.). As the survey results in Table 5 show, the average rating obtained was 4 out of 5, thereby implying that on average there were only a few conversational errors in a typical sized dialog session consisting of about 40 utterances.

Of course, the dataset still contains some noise which is inherited either from the noise in the original catalogs crawled from various websites or because of the process used for creating structured data from unstructured textual descriptions. For example, a product titled “California Bear Logo Flag Republic Flats Bill Snapback” is actually a type of “cap” but when populating the structured data it was wrongly labeled as a “shoe” because “flats” is a valid shoe-type. While such errors exist, they are very minimal and do not affect the overall quality of the dataset. Such noise is expected in any dataset created at this scale. Our manual survey suggested that most dialogs in the dataset have very few (if any) such logical errors.

Tasks

The proposed MMD dataset consists of multimodal, domain-aware conversations between 2 agents, and hence can be used for evaluating a wide variety of tasks. We describe each of these tasks and explain the technical challenges involved:

1. Text Response: Given a context of k turns the task here is to generate the next text response.

2. Image Response: Given a context of k turns the task here is to output the most relevant image(s). There are 2 typical approaches to achieve this:

2.1 Image Retrieval: Given a context of k turns and a database of images, retrieve and rank m images based on their relevance to the given context.

2.2 Image Generation: Given a context of k turns, generate the most relevant image (typically performed using generative models e.g. contextual GANs(Reed et al. 2016; Goodfellow et al. 2014)).

We propose both tasks since the evaluation criteria for each approach is quite different.

3. Employing Domain-knowledge: This is essentially performing tasks (1) and (2) of text and image response

generation using both the unstructured dialog context along with the structured domain knowledge. We propose this as a separate task to evaluate the impact of domain-knowledge.

4. User Modeling: Another important conversation aspect is to study the varying shopping behavior of users e.g. their buying preferences, speed of decision making, etc. Hence, we propose a task to explicitly model the shopper since it impacts the agent’s most appropriate response at each step.

Setup: In this work, we focus on tasks (1) and (2.1) and make two simplifications: (a) We evaluate the text response and image response task separately, which means that the system does not need to decide the modality of the response, and (b) instead of retrieving and ranking all of the images in the catalog/database, the system needs to rank only a given smaller subset of m images, which contain the correct image(s) in addition to a few incorrect ones. This simplified evaluation protocol of “selecting/ranking” the best response was proposed for the Ubuntu Dialog Corpus (Lowe et al. 2015) and helps provide more control on the experimental setup and evaluate individual parts of the system in a more thorough manner.

Dataset Versions

To validate the text generation and image selection tasks in this work, we create two datasets

Version 1: This version includes the “give-image-description” state (as in Table 2) where the system may also provide a short description about the images, while displaying them. For example, in Figure1 the system gives a crisp product description “*michael kors brings you an effortless weekend look with the wrapped espadrille heel*” along with the images

Version 2: This version of the dataset is exactly identical to the first version, except that the utterances corresponding to the “give-image-description” state are missing in this version, i.e. the system will not provide short descriptions of a product, upon displaying its image. The remaining utterances and the flow of the dialogs are identical in both the versions.

It should be noted that the first version poses a harder challenge for the text task, while making the image task somewhat easier, and the second one is more challenging for the image task while being simpler for the text task.

Given this setup, we now propose baseline models for these tasks based on the encode-attend-decode paradigm.

Models

To empirically estimate the feasibility of the tasks described earlier we implement one baseline method (and some variations thereof) for each task, based on the popular hierarchical encode-attend-decode paradigm (Serban et al. 2016a) typically used for (text) conversation systems. We split the description below into two parts (i) Multimodal encoder which is common for the two tasks (ii) Multimodal decoder which is different depending on whether we need to generate a text response or predict an image response.

Mutimodal encoder

As mentioned earlier, for both the tasks, the context contains k utterances where each utterance could either be (i) a text

Type of Complexity	Example State	Example Utterance
Long-Term Context	At the beginning of the dialog the user mentions his budget or size preference and after a few utterances, asks the agent to show something under his budget or size	<i>I like the 4th image. Show me something like it but in style as in this image within my budget.</i>
Quantitative Inference (Counting)	User points to the n th item displayed and asks a question about it	<i>Show me more images of the 3rd product in some different directions</i>
Quantitative Inference (Sorting / Filtering)	User wants sorting/filtering of a list based on a numerical field, e.g. price or product rating	<i>Show me ... within my budget.</i>
Logical Inference	User likes one fashion attribute of the n th image displayed but does not like another attribute of the same	<i>I am keen on seeing something similar to the 1st image but in a different sole material</i>
Visual Inference	System adds a visual description of the product alongside the images	<i>Viscata shoes are lightweight and made of natural jute, premium leather, suedes and woven cloth</i>
Inference over aggregate of Images	User's question can have multiple aspects, drawn from multiple images displayed in the current or past context	<i>List more in the upper material of the 5th image and style as the 3rd and the 5th</i>
Multimodal Inference	User gives partial information in form of images and text in the context	<i>See the first espadrille. I wish to see more like it but in a silver colored type</i>
Inference using domain knowledge and context	Sometimes inferences for the user's questions go beyond the dialog context to understanding the domain	<i>Will the 5th result go well with a large sized messenger bag?</i>
Coreference Resolution / Incomplete Question	Temporal continuity between successive questions from the user may cause some of them to be incomplete or to refer to items or aspects mentioned previously	<i>Show me the 3rd product in some different directions ... What about the product in the 5th image?</i>

Table 4: Anecdotal examples of different aspects of complexity in the MMD dataset. Examples in the 3rd column are snippets of the dialog session illustrated in Figure 1

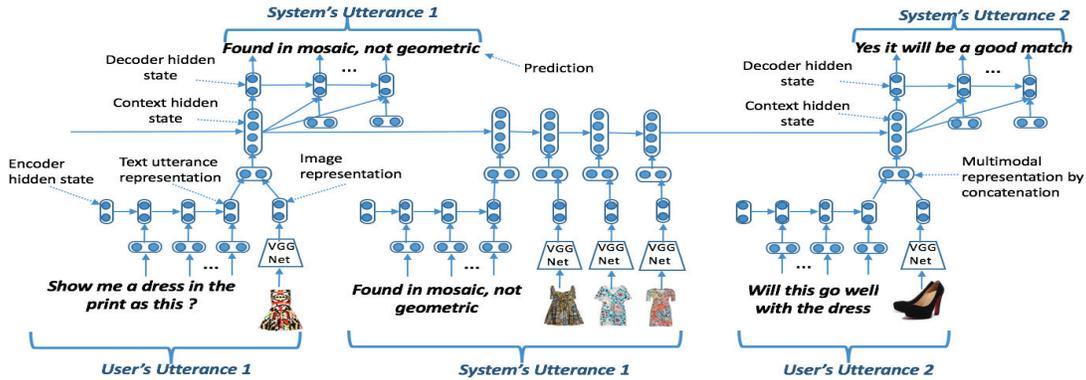


Figure 2: Multimodal Hierarchical Encoder Decoder Architecture for the Text Response Task. The figure shows all multimodal utterances, but in general utterances can have either modality or both.

% of Surveyed 760 Dialogs	Rating	Rating Chart
42.0%	5	dialog is realistic with no errors
30.0%	4	<=2 Conversational Errors
19.2%	3	<=2 Conversational <=2 Logical Errors
6.8%	2	<=2 Conversational <=4 Logical Errors
2.0%	1	>2 Conversational and >4 Logical Errors

Table 5: Domain expert ratings in the qualitative survey.

only or (ii) an image only utterance or (iii) multimodal, containing both text and images (as shown in Figure 2 and Fig. 3). In each of these, we use a multimodal hierarchical encoder for encoding the input representation as below.

(a) Text only utterance: Every text utterance in the context is encoded using a bidirectional RNN network with GRU (Chung et al. 2014) cells in a process similar to the utterance level encoder described in (Serban et al. 2016a). This is the level 1 encoder in the hierarchical encoder.

(b) Image only utterance: If an utterance contains one or more images, we encode each image using a 4096 dimensional representation obtained from the FC6 layer of a VGGNet-16 (Simonyan and Zisserman 2014) convolutional neural network. If an utterance contains multiple images, it is unrolled into a sequence of utterances, with each containing

a single image and/or a single text sequence. This is also a part of the first level in the hierarchical encoder.

(c) Multimodal utterance: The text portion of the multimodal utterance is encoded using the same GRU cells as used for encoding the text only utterance. Similarly, the images in the multimodal utterance are encoded using the same VGGNet-16 as used for the image only utterance. The final representation of the multimodal utterance is simply the concatenation of the individual utterances.

The multimodal utterance representation is then fed to a level two encoder which is again a GRU. This second level (or context-level encoder) essentially encodes the sequence of utterances where the representation of each utterance in the sequence is computed and projected as described above. Fig. 2 and Fig. 3 shows this process of computing the encoder representation for a given multimodal context.

Decoder for generating text responses

As shown in Fig. 2, we use a standard recurrent neural network based decoder with GRU cells. Such a decoder has been used successfully for various natural language generation tasks including text conversation systems (Serban et al. 2016b). We also implemented a version where we couple the decoder with an attention model which learns to attend upon

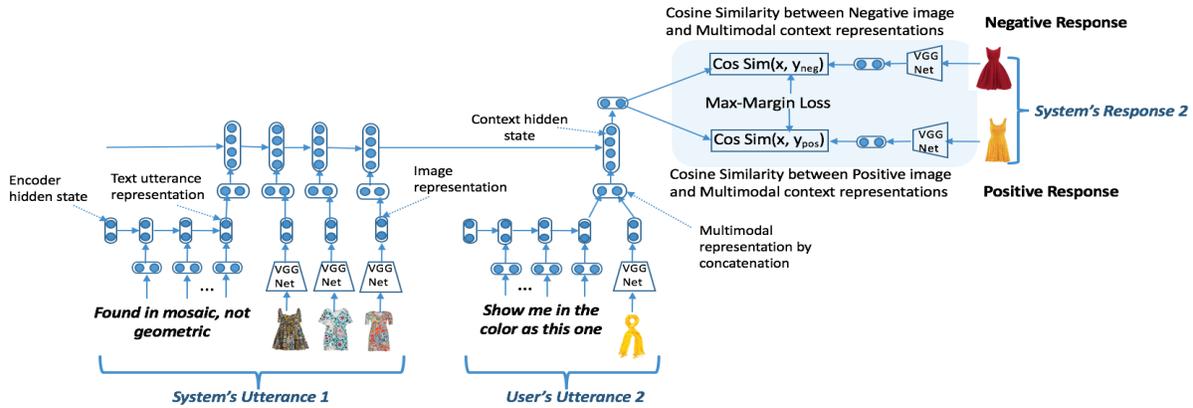


Figure 3: Multimodal Hierarchical Encoder Architecture for Image Response Task. The figure shows a single target positive and negative image, but in general, $m(>1)$ images can be provided as target

different time-steps of the second level encoder (again this has been tried successfully in the context of text conversation systems (Yao et al. 2016)).

Layer for ranking image responses:

The task here is to rank a given set of images depending on their relevance to the context. While training we are given a set of m images for each context of which only $n_{pos,max}$ are relevant for the context. The remaining $m - n_{pos,max}$ are picked from the corresponding false image responses in the dataset. We train the model using a max margin loss. Specifically, we compute the cosine similarity between the learnt image embedding and the encoded representation of the multimodal context. The model is then trained to maximize the margin between the cosine similarity for the correct and the incorrect images. Fig. 3 depicts this for the case when $m = 2$ and $n_{pos,max} = 1$. Due to space constraints we only provide the model's pictorial representations.

Experiments

Now we describe the experimental setup used to evaluate the following models on the two tasks:

- **Hierarchical Encoder Decoder(HRED) (ignoring image context)**, whose architecture is similar to that proposed in (Serban et al. 2016a)
- The proposed **Multimodal Hierarchical Encoder Decoder(HRED)**, (both with and without attention)

Evaluating the Text Response Task

For this task we only considered those dialog turns ending with a text response. The training, validation and test sets sizes are reported in the 6th row of Table 3. We used Adam optimization algorithm and tuned the following hyperparameters using the validation set; learning rate $\in \{1e-3, 4e-4\}$, RNN hidden unit size $\in \{256, 512\}$, text and image embedding size $\in \{256, 512\}$, batch size $\in \{32, 64\}$ and dialog context size $\in \{2, 5, 10\}$. The bracketed numbers indicate the values of each hyperparameter considered. Table 6 summarizes the BLEU and NIST scores used for evaluation.

Ver.	Model (with Context Size)	Text Task		Image Task ($m=5$)		
		Bleu	Nist	R@1	R@2	R@3
V1	Text-Only HRED (5)	14.58	2.61	0.46	0.64	0.75
	Multimodal HRED (2)	20.42	3.09	0.72	0.86	0.92
	Multimodal HRED (5)	19.73	2.94	0.71	0.86	0.92
	Attention based Multimodal HRED (2)	20.2	2.96	0.71	0.86	0.92
	Attention based Multimodal HRED (5)	19.37	2.51	0.68	0.84	0.91
V2	Text-Only HRED (5)	35.9	5.14	0.44	0.6	0.72
	Multimodal HRED (2)	56.67	7.51	0.69	0.85	0.90
	Multimodal HRED (5)	56.15	7.27	0.68	0.84	0.90
	Attention based Multimodal HRED (2)	56.74	7.65	0.68	0.84	0.90
	Attention based Multimodal HRED (5)	54.58	6.92	0.66	0.83	0.89

Table 6: Performance of the different models on the ‘‘Text Response Generation’’ and ‘‘Best Image Selection’’ Task.

Evaluating Image Response Task

During training and evaluation for this task we only consider those dialog turns ending in an image response from the system. The training, validation and test sets sizes are reported in the 5th row of Table 3. Both during training and testing, the model is provided with $m=5$ target images out of which only $n_{pos,max}=1$ is relevant and at test time the model has to rank the images in order of their relevance as a response to the given context. The hyperparameters of the model were tuned in the same way as mentioned above. Note that for evaluating the image response in selection/ranking mode, a system would also need negative training and test examples alongside the correct ones. Negative examples are generated by either sampling an item from the wrong fashion category (e.g. shoe in place of a bag) or a wrong sub-category of the

Id	Ver	Text Task		Image Task					
				m=50			m=100		
		Bleu	Nist	r@1	r@2	r@3	r@1	r@2	r@3
10	V1	48.5	6.42	-	-	-	-	-	-
	V2	47.8	6.31	-	-	-	-	-	-
14	V1	10.5	0.78	-	-	-	-	-	-
	V2	14.3	1.18	-	-	-	-	-	-
9	V1	29.0	1.42	-	-	-	-	-	-
	V2	33.0	2.7	-	-	-	-	-	-
11	V1	67.2	6.23	-	-	-	-	-	-
	V2	70.4	6.42	-	-	-	-	-	-
6	V1	4.65	0.73	0.18	0.32	0.43	0.11	0.20	0.27
	V2	94.0	8.97	0.18	0.30	0.41	0.11	0.19	0.26
13	V1	27.5	1.88	0.24	0.38	0.49	0.15	0.25	0.34
	V2	98.1	4.03	0.22	0.36	0.48	0.13	0.22	0.30
7	V1	100	6.71	0.26	0.40	0.50	0.13	0.23	0.31
	V2	100	6.71	0.23	0.37	0.47	0.15	0.25	0.32
8	V1	100	2.32	0.23	0.37	0.48	0.15	0.25	0.33
	V2	100	2.32	0.21	0.35	0.46	0.13	0.23	0.31
12	V1	100	5.42	0.24	0.38	0.49	0.09	0.18	0.26
	V2	100	5.42	0.22	0.37	0.47	0.14	0.24	0.32

Table 7: Best Model’s performance on dialog states described in Table 2 (‘Id’ refers to the ID of the state-type in Table 2) and V1, V2 refer to the two versions of the dataset, m refers to the size of target image set (one correct, rest incorrect) to be ranked by the model and is varied from 50 to 100

target item (e.g a backpack in place of a sachel bag) or items violating certain attribute criteria provided by the user.

We use $Recall@top-m$ as the evaluation metrics where $top-m$ is varied from 1 to 3, and the model prediction is considered to be correct only if the true response is among the $top-m$ entries in the ranked list. These are summarized in Table 6.

Discussions

We make a few observations from the results.

- For both tasks, the “Multimodal HRED model with image sequence” performs significantly better than both the unimodal baseline HRED models thus suggesting that adding images indeed improves inference for both the tasks.
- Comparing the performance of the text response tasks on the two dataset versions, it is obvious that the model performs fairly well on all kinds of text responses except the “give-image-description” response, which in itself is a very hard task as it exploits both visual features and domain knowledge or other catalog information.
- Further, comparing the image response performance for the two dataset version, we observe that having additional textual descriptions of the images in a dialog context can help in better image response selection, which is intuitive.

- Adding attention does not improve performance. Though counter intuitive, this suggests the need for better multimodal attention models.
- In Table 7, we have reported the performance of the best model trained for the Text and Image Response Task (as per Table 6) on the 9 most significant and frequently occurring user-initiated dialog states described earlier in Table 2. As is evident from the table, the text response performance shows a high variance over the dialog states especially for Dataset V1, thus indicating that wherever the system needs to respond with a short product description, requiring core domain knowledge, it performs poorly.
- Further, in Table 7, we also report the performance on the Image Task obtained by varying m . When we use 50 or 100 candidate images (instead of 5) we see a sharp decline in the performance, indicating that a better model is needed to retrieve and rank images from a large corpus of images.
- Overall, we feel there is enough scope for improvement and the current models only establish the feasibility of the two tasks. Further, we believe that benchmarking the performance on salient dialog states will allow for more focused research by understanding the challenges and complexities involved in each of these states.

To facilitate further research on multimodal systems, the MMD dataset created as a part of this work will be made available at <https://github.com/iitm-nlp-miteshk/AmritaSaha/tree/master/MMD> (please copy paste the URL in a browser instead of clicking on it). This URL will contain the following resources:

- the train, valid, test splits of the two versions of the MMD dataset and the script to extract the state-wise data for each of the states elaborated in Table 2
- the multimodal catalog data in the raw form (before parsing) consisting of unstructured text descriptions and images of the product. But it should be noted that for the current benchmarked models, we only used the image information from the catalog and the multimodal context from the ongoing dialog
- the domain specific knowledge-base (*i.e.*, the fashion taxonomy, attribute lexicons, style-tips, celebrity profiles *etc.*) curated from the parsed catalog

Conclusion

In this paper, we introduced the Multimodal Dialogs (MMD) dataset curated by working closely with a group of 20 fashion retail experts and consisting of over 150K multimodal conversation sessions between shoppers and sales agents. We proposed 5 new sub-tasks along with their evaluation methodologies. We also showcased 2 multimodal neural baselines using the encode-attend-decode paradigm and demonstrated their performance on both text response generation and best image response selection. Their performance demonstrate the feasibility of the involved sub-tasks and highlight the challenges present. Finally, we suggest new research directions for addressing the challenges in multimodal conversation.

References

- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2425–2433.
- Bordes, A., and Weston, J. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683.
- Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2016. Visual dialog. *CoRR* abs/1611.08669.
- de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. C. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. *CoRR* abs/1611.08481.
- Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2672–2680.
- Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, 2-4 September 2015, Prague, Czech Republic*, 285–294.
- Maharaj, T.; Ballas, N.; Courville, A. C.; and Pal, C. J. 2016. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *CoRR* abs/1611.07810.
- Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G. P.; and Vanderwende, L. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR* abs/1701.08251.
- Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*.
- Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, 172–180.
- Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, 3776–3783. AAAI Press.
- Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. *CoRR* abs/1605.06069.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.
- Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2048–2057.
- Yao, K.; Peng, B.; Zweig, G.; and Wong, K. 2016. An attentional neural conversation model with improved specificity. *CoRR* abs/1606.01292.
- Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 4584–4593.
- Yu-I Ha, Sejeong Kwon, M. C. J. J. 2017. Fashion conversation data on instagram. *CoRR* abs/1704.04137.
- Zeng, K.; Chen, T.; Chuang, C.; Liao, Y.; Niebles, J. C.; and Sun, M. 2016. Leveraging video descriptions to learn video question answering. *CoRR* abs/1611.04021.