Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System

Hardik Chauhan[1*], **Mauajama Firdaus**[2*], **Asif Ekbal**[2], **Pushpak Bhattacharyya**[2]
[1] Department of Electrical Engineering, Indian Institute of Technology Roorkee, India
[2] Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
(chauhanhardik23@gmail.com),(mauajama.pcs16,asif,pb)@iitp.ac.in

Abstract

Multimodal dialogue systems have opened new frontiers in the traditional goal-oriented dialogue systems. The state-of-the-art dialogue systems are primarily based on unimodal sources, predominantly the text, and hence cannot capture the information present in the other sources such as videos, audios, images etc. With the availability of large scale multimodal dialogue dataset (MMD) (Saha et al., 2018) on the fashion domain, the visual appearance of the products is essential for understanding the intention of the user. Without capturing the information from both the text and image, the system will be incapable of generating correct and desirable responses. In this paper, we propose a novel position and attribute aware attention mechanism to learn enhanced image representation conditioned on the user utterance. Our evaluation shows that the proposed model can generate appropriate responses while preserving the position and attribute information. Experimental results also prove that our proposed approach attains superior performance compared to the baseline models, and outperforms the state-of-the-art approaches on text similarity based evaluation metrics.

1 Introduction

With the advancement in Artificial Intelligence (AI), dialogue systems have become a prominent part in today's virtual assistant, which helps users to converse naturally with the system for effective task completion. Dialogue systems focus on two broad categories - open domain conversations with casual chit chat and goal-oriented systems where the system is designed to solve a particular task for the user belonging to a specific domain. Response generation is a crucial component of every conversational agent. The task of "how to say" the information to the user is the primary objective of every response generation module. One of the running goals of AI is to bring language and vision together in building robust dialogue systems. Advances in visual question answering (VQA) (Kim et al., 2016; Xiong et al., 2016; Ben-Younes et al., 2017), and image captioning (Anderson et al., 2018; Chen et al., 2018) have ensured interdisciplinary research in natural language processing (NLP) and computer vision. Recently, several works in dialogue systems incorporating both vision and language (Das et al., 2017a; Mostafazadeh et al., 2017) have shown promising research directions.

Goal oriented dialogue systems are majorly based on textual data (unimodal source). With increasing demands in the domains like retail, travel, entertainment, conversational agents that can converse by combining different modalities is an essential requirement for building the robust systems. Knowledge from different modalities carries complementary information about the various aspects of a product, event or activity of interest. By combining information from different modalities to learn better representation is crucial for creating robust dialogue systems. In a multimodal setup, the provision of different modalities assists both the user and the agent in achieving the desired goal. Our work is established upon the recently proposed Multimodal Dialogue (MMD) dataset (Saha et al., 2018), consisting of e-commerce (fashion domain) related conversations. The work focused on generating textual responses conditioned on the conversational history consisting of both text and image.

In the existing task-oriented dialogue systems, the inclusion of visually grounded dialogues- as in the case of MMD dataset- has provided exciting new challenges in the field of interactive dialogue systems. In contrast to VQA, multimodal

[*] First two authors are jointly the first authors

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5437–5447
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

dialogues have conversations with more extended contextual dependency, and a clear end-goal. As opposed to a static image in VQA, MMD deals with dynamic images making the task even more challenging. In comparison to the previous slot-filling dialogue systems on textual data (Young et al., 2013; Rieser and Lemon, 2011), MMD provides an additional visual modality to drive the conversation forward.

In this work, we propose an entirely data-driven response generation model in a multi-modal setup by combining the modalities of text and images. In Figure 1, we present an example from the MMD dataset. It is a conversation between the user and the system in a multimodal setting on the fashion domain. From the example, it is understood that the position of images is essential for the system to fulfill the demands of the user. For example, in figure, the U3 utterance *"Can you tell me the type of colour in the 1st image"* needs position information of the particular image from the given set of images. To handle such situations, we incorporate position embeddings to capture ordered visual information. The underlying motivation was to capture the correct image information from the text; hence, we use position aware attention mechanism. From Figure 1, in utterance U5, we can see that the user is keen on different aspects of the image as well. In this case, user is interested in the *"print as in the 2nd image"*. To focus and capture the different attributes from the image representation being considered in the text, we apply attribute aware attention on the image representation. Hence in order to handle such situations present in the dataset, we apply both position and attribute aware attention mechanisms to capture intricate details from the image and textual features. For effective interaction among the modalities, we use Multimodal Factorized Bilinear (MFB) (Yu et al., 2017) pooling mechanism. Since multimodal feature distribution varies dramatically, hence the integrated image-text representations obtained by such linear models may not be sufficient in capturing the complex interactions between the visual and textual modalities. The information of the present utterance, image and the contextual history are essential for better response generation (Serban et al., 2015).

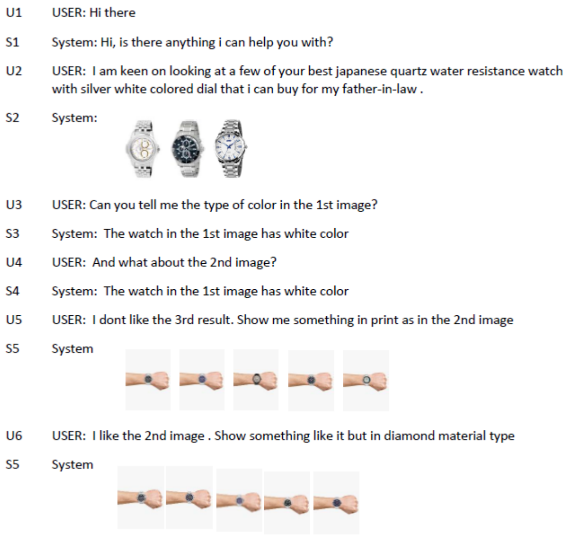
The key contributions/highlights of our current work are as follows:

- We employ a position-aware attention mechanism to incorporate the ordered visual information and attribute-aware attention mechanism to focus on image conditioned on the attributes discussed in the text.

- We utilize Multi-modal Factorized Bilinear (MFB) model to fuse the contextual information along with image and utterance representation.

- We achieve state-of-the-art performance for the textual response generation task on the MMD dataset.

The rest of the paper is structured as follows: In section 2, we discuss the related works. In Section 3, we explain the proposed methodology followed by the dataset description in Section 4. Experimental details and evaluation metrics are reported in Section 5. Results along with necessary analysis are presented in Section 6. In Section 7, we conclude the paper along with future research direction.



U1 USER: Hi there

S1 System: Hi, is there anything i can help you with?

U2 USER: I am keen on looking at a few of your best japanese quartz water resistance watch with silver white colored dial that i can buy for my father-in-law .

S2 System:

U3 USER: Can you tell me the type of color in the 1st image?

S3 System: The watch in the 1st image has white color

U4 USER: And what about the 2nd image?

S4 System: The watch in the 1st image has white color

U5 USER: I dont like the 3rd result. Show me something in print as in the 2nd image

S5 System

U6 USER: I like the 2nd image . Show something like it but in diamond material type

S5 System

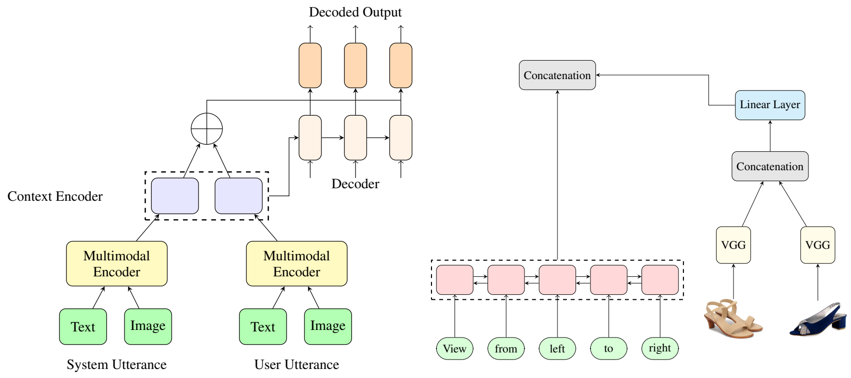
Figure 1: An example from the MMD dataset

2 Related Work

Research on dialog systems have been a major attraction since a long time. In this section we briefly discuss some of the prominent research carried out on single and multi-modal dialog systems.

2.1 Unimodal Dialogue Systems

Dialogue systems have mostly focused on single modal source such as text. Hence, there have been



(a) Overall model architecture with Multimodal encoder followed by context encoder and the decoder module

(b) Multimodal encoder with simple concatenation of text and image representations

Figure 2: Block Diagram of the MHRED model; Left image is the overall system architecture for text generation; Right image is the baseline encoder model

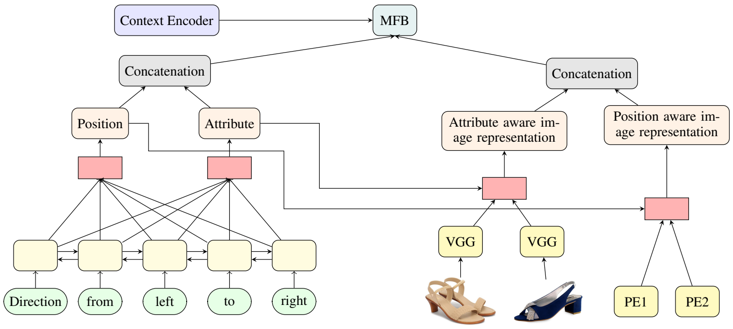


Figure 3: Proposed Multimodal Encoder with Position and Attribute aware Attention with MFB fusion

several works carried out on data-driven textual response generation. To help the users achieve their desired goals, response generation provides the medium through which a conversational agent can communicate with its user. In (Ritter et al., 2011), the authors used social media data for response generation following the machine translation approach. The effectiveness of deep learning has shown remarkable improvement in dialogue generation. Deep neural models have been quite beneficial for modelling conversations in (Vinyals and Le, 2015; Li et al., 2016a,b; Shang et al., 2015). A context-sensitive neural language model was proposed in (Sordoni et al., 2015), where the model chooses the most probable response given the textual conversational history. In (Serban et al., 2015, 2017), the authors have proposed a hierarchical encoder-decoder model for capturing the dependencies in the utterances of a dialogue. Conditional auto-encoders have been employed in (Zhao et al.; Shen et al., 2018) that generate diverse replies by capturing discourse-level information in the encoder. Our current work differentiates from these existing works in dialogue systems in a way that we generate the appropriate responses by capturing information from both the text and image, conditioned on the conversational history.

2.2 Multimodal Dialogue Systems

With the recent shift in interdisciplinary research, dialogue systems combining different modalities (text, images, video) have been investigated for creating robust conversational agents. Dialogue generation combining information from text and images (Das et al., 2017a,b; Mostafazadeh et al., 2017; Gan et al., 2019; De Vries et al., 2017) has been successful in bridging the gap between vision and language. Our work differs from these as the conversation in Multimodal Dialogue (MMD) dataset (Saha et al., 2018) deals with multiple images and the growth in conversation is dependent on both image and text as opposed to a conversation with a single image. Lately, with the release of DSTC7 dataset, video and textual modalities have been explored in (Lin et al., 2019; Le et al., 2019). Prior works on MMD dataset reported in (Agarwal et al., 2018b,a; Liao et al., 2018) have captured the information in the form of knowledge bases using hierarchical encoder-decoder model.

Our work is different from these existing works on MMD dataset in the sense that we incorporate position and attribute aware attention mechanism for capturing ordered information and minute details such as colour, style etc. from the image representations for more accurate response generation. Our method, unlike the previous works, make use of the MFB technique for better information fusion across different modalities. The approach that we propose to capture and integrate information from image and text is novel. We successfully demonstrate the effectiveness of our proposed model in generating responses through sufficient empirical analysis.

3 Methodology

In this section we firstly define the problem and then present the details of the proposed method.

3.1 Problem Definition

In this paper, we address the task of textual response generation conditioned on conversational history as proposed in (Saha et al., 2018). The dialogue consists of text utterances along with multiple images and given a context of k turns the task here is to generate the next text response. More precisely, given an user utterance $U_k = (w_{k,1}, w_{k,2}, ..., w_{k,n})$, a set of images $I_k = (img_{k,1}, img_{k,2}, ..., img_{k,n'})$ and a conversational history $H_k = ((U_1, I_1), (U_2, I_2), ..., (U_{k-1}, I_{k-1}))$ the task is to generate the next textual response $Y_k = (y_{k,1}, y_{k,2}, ..., y_{k,n''})$.

3.2 Hierarchical Encoder Decoder

We construct a response generation model, as shown in Figure 2(a), which is an extension of the recently introduced Hierarchical Encoder Decoder (HRED) architecture (Serban et al., 2016, 2017). As opposed to the standard sequence to sequence models (Cho et al., 2014; Sutskever et al., 2014), the dialogue context is modelled by a separate context Recurrent Neural Network (RNN) over the encoder RNN, thus forming a hierarchical encoder. The multimodal HRED (MHRED) is built upon the HRED to include text and image modalities. The key components of MHRED are the utterance encoder, image encoder, context encoder and decoder.

Utterance Encoder: Given an utterance U_m, a bidirectional Gated Recurrent Units (BiGRU) (Bahdanau et al., 2014) is employed to encode each word $w_{m,i},\ i \in (1, ..., n)$ represented by d-dimensional embeddings into the hidden vectors $h_{m,U,i}$.

$$\overrightarrow{h_{U,m,i}} = GRU_{u,f}(w_{m,i}, \overrightarrow{h_{U,m,i-1}}) \qquad (1)$$

$$\overleftarrow{h_{U,m,i}} = GRU_{u,b}(w_{m,i}, \overleftarrow{h_{U,m,i-1}}) \qquad (2)$$

$$h_{U,m,i} = [\overrightarrow{h_{U,m,i}}, \overleftarrow{h_{U,m,i}}] \qquad (3)$$

Image Encoder: A pre-trained VGG-19 model (Simonyan and Zisserman, 2014) is used to extract image features for all the images in a given dialogue turn. The concatenation of single image features is given as input to a single linear layer to obtain a global image context representation.

$$F_{m,i} = VGG(img_{m,i}) \qquad (4)$$

$$F_m = Concat(F_{k,1}, F_{k,2}, ..., F_{k,n'}) \qquad (5)$$

$$h_{I,m} = ReLU(W_I F_m + b_I) \qquad (6)$$

where W_I and b_I are the trainable weight matrix and biases, respectively. The number of images in a single turn is ≤ 5; hence, zero vectors are considered in the absence of images.

Context-level Encoder: The final hidden representations from both image as well as text encoders are concatenated for every turn and are fed as input to the context GRU, as shown in Figure 2(b). A hierarchical encoder is built on top of the image and text encoder to model the dialogue history. The final hidden state of the context GRU serves as the initial state of the decoder GRU.

$$h_{c,m} = GRU_c([h_{I,m}; h_{U,m,n}], h_{c,m-1}) \qquad (7)$$

Decoder: In the decoding stage, the decoder is another GRU that generates words sequentially conditioned on the final hidden state of the context GRU and the previously decoded words. Attention mechanism similar to (Luong et al., 2015) is incorporated to enhance the performance of the decoder GRU. The attention layer is applied to the hidden state of context encoder using decoder state d_t as the query vector. The concatenation of the context vector and the decoder state is used to compute a final probability distribution over the output tokens.

$$h_{d,t} = GRU_d(y_{k,t-1}, h_{d,t-1}) \qquad (8)$$

$$\alpha_{t,m} = softmax(h_{c,m}^T W h_{d,t}) \tag{9}$$

$$c_t = \sum_{m=1}^{k} \alpha_{t,m} h_{c,m}, \tag{10}$$

$$\tilde{h}_t = tanh(W_{\tilde{h}}[h_{d,t}; c_t]) \tag{11}$$

$$P(y_t/y_{<t}) = softmax(W_V \tilde{h}_t) \tag{12}$$

where, W_h, W_V and $W_{\tilde{h}}$ are trainable weight matrices.

3.3 Proposed Model

To improve the performance of the MHRED model, rather than just concatenating the representations of the text and image encoder we apply an attention layer to mask out the irrelevant information. In our case, we apply attention to learn where to focus and what to focus upon as described in the user utterance. To decouple these two tasks we augment the encoder with position and attribute aware attention mechanisms.

Position-aware Attention: In the baseline MHRED model, we incorporate position information of the images to improve the performance of the system. For example, *"List more in colour as the 4th image and style as in the 1st image"*, the ordered information of the images is essential for the correct textual response by the agent to satisfy the needs of the user. Hence, the knowledge of every image with respect to its position is necessary so that the agent can capture the information and fulfill the objective of the customer. The lack of position information of the images in the baseline MHRED model causes quite a few errors in focusing on the right image. To alleviate this issue, we fuse position embedding of every image with the corresponding image features. The position of every image is represented by position embedding PE_i, where, $PE = [PE_1, ..., PE_{n'}]$. This information is concatenated to the corresponding image features. To compute self attention (Wang et al., 2017) we represent textual features as $H_U = [h_{U,1}, ..., h_{U,n}]$.

$$\alpha_p = softmax({W_p}^T H_U), U_p = \alpha_p {H_U}^T \tag{13}$$

We use the self-attended text embedding as a query vector U_p to calculate the attention distribution over the position embedding PE.

$$\beta_p = softmax({U_p}^T W_{p'} PE), I_p = \beta_p PE^T \tag{14}$$

where, ${W_p}^T$ and $W_{p'}$ are trainable parameters.

Attribute-aware Attention: To focus on different attributes of the image mentioned in the text, we employ attribute-aware attention.

$$\alpha_a = softmax({W_a}^T H_U), U_a = \alpha_a {H_U}^T \tag{15}$$

The self-attended text embedding is used as query vector U_a to compute the attention distribution over the image feature represented by $H_I = [h_{I,1}, ..., H_{I,n'}]$.

$$\beta_a = softmax({U_a}^T W_{a'} H_I), I_a = \beta_a {H_I}^T \tag{16}$$

where, ${W_a}^T$ and $W_{a'}$ are trainable parameters.

Finally, in our proposed model, as shown in Figure 3, we incorporate position-aware and attribute-aware attention mechanisms to provide focused information conditioned on the text utterance. We concatenate U_a and U_p vectors for the final utterance representations U_f, I_a and I_p vectors as the final image representation I_f. The output of the context encoder h_c along with I_f and U_f serves as input to the MFB module. Here, we compute the MFB between I_f and U_f.

$$z = SumPooling(W_m {U_f}^T \circ W_{m'} {I_f}^T, k') \tag{17}$$

$$z = sign(z)|z|^{0.5}, z = z^T/||z|| \tag{18}$$

where, W_m and $W_{m'}$ are the trainable parameters, and $SumPooling$ function is same as described in (Gan et al., 2019). Similarly, we take a pairwise combination of I_f, U_f and h_c as the final output of our multimodal fusion module. Hence, the final multimodal fusion can be represented by $h_d = [MFB(U_f, I_f), MFB(U_f, h_c), MFB(I_f, h_c)]$, where h_d is used to initialize the decoder.

3.4 Training and Inference

We employ commonly used teacher forcing (Williams and Zipser, 1989) algorithm at every decoding step to minimize negative log-likelihood on the model distribution. We define $y^* = \{y_1^*, y_2^*, \ldots, y_m^*\}$ as the ground-truth output sequence for a given input

$$\mathcal{L}_{ml} = -\sum_{t=1}^{m} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*) \tag{19}$$

We apply uniform label smoothing(Szegedy et al., 2016) to alleviate the common issue of low diversity in dialogue systems, as suggested in (Jiang and de Rijke, 2018).

3.5 Baseline Models

For our experiment, we develop the following models:

Model 1 (MHRED): The first model is the baseline MHRED model described in Section 3.2.

Model 2 (MHRED + A): In this model, we apply attention (A) on the text and image features rather than merely concatenating the features.

Model 3 (MHRED + A + PE): In this model, position embeddings (PE) of every image is concatenated with the respective image features to provide ordered visual information of the images.

Model 4 (MHRED + PA): Self-attention on the text representations with respect to position information is computed to generate a query vector. This query vector is used to learn the attention distribution on the position embeddings to focus on the discussed image in user utterance.

Model 5 (MHRED + AA): To learn the different attributes discussed in the text we apply self-attention on the text representation and compute a query vector that attends the image representation in accordance to the attributes in the text.

Model 6 (MHRED + PA + AA): In this model, the final text and image representations, denoted as U_f and I_f, respectively, and obtained after applying the position and attribute aware attention, are concatenated and fed as input to the context encoder.

Model 7 (MHRED + MFB(I, T)): MFB module is employed to learn the complex association between the textual and visual features. The final text representation (T) U_f and the final image representation (I) I_f are fed as input to the MFB module.

Model 8 (MHRED + MFB(I,T,C)): In this model, we concatenate the pairwise output of the MFB module on the contextual information (C), that is the output of context encoder $h_{c,i}$ along with text and image representations.

4 Datasets

Our work is built upon the Multimodal Dialogue (MMD) dataset (Saha et al., 2018). The MMD dataset comprises of 150k chat sessions between the customer and sales agent. Table 1 lists the detailed information about the MMD dataset. Domain-specific knowledge in the fashion domain was captured during the series of customer-agent interactions. The dialogues incorporate text and image information seamlessly in a conversation bringing together multiple modalities for creating advanced dialogue systems. The dataset poses new challenges for multimodal, goal-oriented dialogue containing complex user utterances. For example, *"Can you show me the 5th image in different orientations within my budget?"*, requires quantitative inference such as *filtering, counting and sorting*. Bringing the textual and image modalities together, multimodal inference makes the task of generation even more challenging, for example, *"See the second stilettos, I want to see more like it but in a different colour"*. In our work, we use a different version of the dataset as described in (Agarwal et al., 2018a,b) to capture the multiple images, in turn, as one concatenated context vector for every turn in a given dialogue.

Dataset Statistics	Train	Valid	Test
Number of dialogues	105,439	22,595	22,595
Avg. turns per Dialogue	40	40	40
No. of Utterances with Image Response	904K	194K	193K
No. of Utterances with Text Response	1.54M	331K	330K
Avg. words in Text Response	14	14	14

Table 1: Dataset statistics of MMD

5 Experiments

In this section we present the implementation details and the evaluation metrics (automatic and human) that we use for measuring the model performance.

5.1 Implementation Details

All the implementations are done using the PyTorch[1] framework. We use 512-dimensional word embedding and 10-dimensional position embedding as described in (Vaswani et al., 2017). We use the dropout(Srivastava et al., 2014) with probability 0.45. During decoding, we use a beam search with beam size 10. We initialize the model parameters randomly using a Gaussian distribution with Xavier scheme (Glorot and Bengio, 2010). The hidden size for all the layers is 512. We employ AMSGrad (Reddi et al., 2019) as the optimizer for model training to mitigate the slow convergence issues. We use uniform label smoothing with $\epsilon = 0.1$ and perform gradient clipping when gradient norm is over 5. For image representation,

[1] https://pytorch.org/

FC6(4096 dimension) layer representation of the VGG-19 (Simonyan and Zisserman, 2014), pre-trained on ImageNet is used.

5.2 Automatic Evaluation

For evaluating the model we report the standard metrics like BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Lavie and Agarwal, 2007) employing the evaluation scripts made available by (Sharma et al., 2017).

5.3 Human Evaluation

To understand the quality of responses, we adopt human evaluation to compare the performance of different models. We randomly sample 700 responses from the test set for human evaluation. Given an utterance, image along with the conversation history were presented to three human annotators, with post-graduate level of exposure. They were asked to measure the correctness and relevance of the responses generated by the different models with respect to the following three metrics:

1. Fluency (F): The generated response is grammatically correct and is free of any errors. **2.** Relevance (R): The generated response is in accordance to the aspect being discussed (style, colour, material, etc.), and contains the information with respect to the conversational history. Also, there is no loss of attributes/information in the generated response.

We follow the scoring scheme for fluency and relevance as- 0: incorrect or incomplete, 1: moderately correct, and 2: correct. We compute the Fleiss' kappa (Fleiss, 1971) for the above metrics to measure inter-rater consistency. The kappa score for fluency is 0.75 and relevance is 0.77 indicating "substantial agreement".

6 Results and Analysis

In this section we present the detailed experimental results using automatic and human evaluation metrics both. In addition we also report the errors that our current model encounters.

6.1 Automatic Evaluation Results

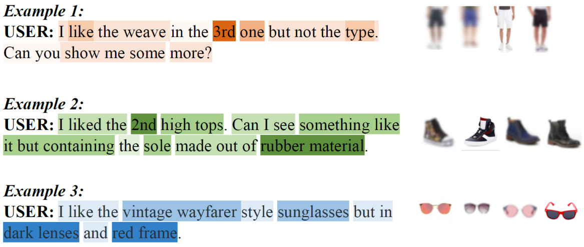
Results of the different models are presented in Table 2. The proposed model performs better than the other baselines for all the evaluation metrics, and we find this improvement to be statistically significant [2]. The results are reported for context size 5 due to its superior performance in comparison to the context size 2, as shown in (Agarwal et al., 2018a,b). The MHRED model is a decent baseline with good scores (0.6725 ROUGE-L, 0.4454 BLEU). The application of attention over the text and image representations, as opposed to the concatenation, provides an absolute improvement of (+0.85%) in METEOR as well as in the other metrics. To give the ordered visual information in Model 3, we incorporate positional embedding for the images which boost the performance of text generation by (+0.94%) in BLEU score and (+0.58%) in ROUGE-L.

Description	Model	BLEU 4	METEOR	ROUGE L
State-of-the-arts	*MHRED-attn (Agarwal et al., 2018a)*	0.4451	0.3371	0.6799
	MHRED-attn-kb (Agarwal et al., 2018b)	0.4634	0.3480	0.6923
Baseline Models	*MHRED*	0.4454	0.3367	0.6725
	MHRED + A	0.4512	0.3452	0.6754
	MHRED + A + PE	0.4548	0.3476	0.6783
	MHRED + PA	0.4781	0.3521	0.7055
	MHRED + AA	0.4763	0.3511	0.7063
	MHRED + PA + AA	0.4810	0.3569	0.7123
	MHRED + MFB(I,T)	0.4791	0.3523	0.7115
	MHRED + MFB(I,T,C)	0.4836	0.3575	0.7167
Our Proposed Model	***MHRED + PA + AA + MFB(I,T)***	**0.4928**	**0.3689**	**0.7211**
	MHRED + PA + AA + MFB(I,T,C)	**0.4957**	**0.3714**	**0.7254**

Table 2: Results of different models on MMD dataset. Here, A: Attention, PE: Positional embeddings, PA: Position-aware attention, AA: Attribute-aware attention, MFB (I,T): MFB fusion on image (I) and text (T) representations, MFB(I,T,C): MFB fusion on I,T and context (C)

The improved performance shows the effectiveness of position embedding for the images in a multimodal dialogue setting. The efficiency of position-aware and attribute-aware attention mechanism (Model 6) can be seen in the increased performance of the model with respect to Model 4 and Model 5 with an improvement of 0.68% and 0.6% in ROUGE-L metric, respectively. The MFB based fusion technique helps to improve the performance of the generation model (Model 8) with an improvement of 3.82% in BLEU score with respect to the baseline model, whereas it shows 0.26% improvement in BLEU score in comparison to Model 6. The final proposed model (MHRED + PA + AA + MFB(I,T,C)) after incorporating the position and attribute aware attention mechanisms along with MFB fusion attains the state-of-the-art performance with an improvement of 3.23% in BLEU score, 3.31% in ROUGE-L and 2.34% in METEOR in comparison to the existing approaches (Agarwal et al., 2018b).

[2] we perform statistical significance t-test (Welch, 1947) and it is conducted at 5% (0.05) significance level



Example 1:
USER: I like the weave in the 3rd one but not the type. Can you show me some more?

Example 2:
USER: I liked the 2nd high tops. Can I see something like it but containing the sole made out of rubber material.

Example 3:
USER: I like the vintage wayfarer style sunglasses but in dark lenses and red frame.

Figure 4: Position and Attribute aware Attention Visualization

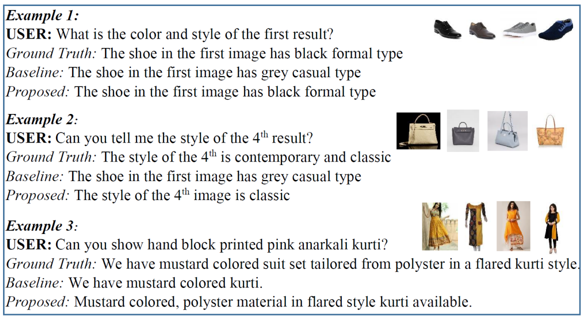
In Figure 4, we show the attention visualization to demonstrate the effectiveness of our proposed position and attribute aware attention mechanisms. Example 1 in the figure shows that the model can focus on the correct image (in this case, the 3rd image) with the help of position-aware attention mechanism as the focus is given to the word *3rd* in the utterance. Example 2 shows the effect of both position and attribute aware attention mechanism that helps in more accurate response generation. The positional word *2nd* along with the attribute *rubber* has obtained maximum focus in the given example. While in Example 3, we can see the effect of attribute aware attention mechanism with maximum attention given to the keywords such as *dark, red, frame* in the utterance.

6.2 Human Evaluation Results

In Table 3, we present the evaluation results of human. In case of fluency, the baseline MHRED model and the proposed model have shown quite similar performance. While for the relevance metric our proposed model has shown better performance with an improvement of 7.47% in generating the correct responses. This may be due the reason that our proposed model focuses on the relevant information in the text as well as the image, and generate more accurate and informative responses. All the results are statistically significant as we perform Welch's t-test (Welch, 1947) and it is conducted at 5% (0.05) significance level.

Description	**Model**	**Fluency**			**Relevance**		
		0	1	2	0	1	2
Baseline	*MHRED*	18.64	39.66	41.70	13.41	39.83	46.76
Proposed	*MHRED + PA + AA + MFB(I,T,C)*	15.54	42.71	41.75	7.36	38.14	54.23

Table 3: Human evaluation results for Fluency and Relevance (All values are in percentages.)



Example 1:
USER: What is the color and style of the first result?
Ground Truth: The shoe in the first image has black formal type
Baseline: The shoe in the first image has grey casual type
Proposed: The shoe in the first image has black formal type

Example 2*:*
USER: Can you tell me the style of the 4th result?
Ground Truth: The style of the 4th is contemporary and classic
Baseline: The shoe in the first image has grey casual type
Proposed: The style of the 4th image is classic

Example 3*:*
USER: Can you show hand block printed pink anarkali kurti?
Ground Truth: We have mustard colored suit set tailored from polyster in a flared kurti style.
Baseline: We have mustard colored kurti.
Proposed: Mustard colored, polyster material in flared style kurti available.

Figure 5: Examples of Responses Generated by the Different Models

6.3 Error Analysis

We analyse the outputs generated from our proposed model to perform a detailed qualitative analysis of the responses. In Figure 5, we present a few examples of the responses generated by the different models given the image and utterance as an input. Some commonly occurring errors include:

1. Unknown tokens: As the baseline MHRED model uses the basic sequence to sequence framework, the number of unknown tokens is predicted the most in this case. The model also often predicts 'end of sequence' token just after the 'out of vocabulary' token, thus leaving sequences incomplete. Gold: *..the type of the chinos is cargo in the 1st and 2nd image*; Predicted: .. *the type*

2. Extra information: The proposed model sometimes generates extra informative sentences than in the ground-truth response due to multiple occurrences of these attributes together in the data: Gold: *the jackets in the 1st, 2nd and 5th images will suit well for dry clean*; Predicted: *the jackets in the 1st, 2nd and 5th images will suit well for dry clean, regular, cold, hand clean.*

3. Repetition: The baseline, as well as the proposed model in a few cases, go on repeating the information present in a given utterance: Gold: *it can go well with cropped type navy sweater*; Predicted: *it can go well with navy style, navy neck, navy style, navy neck sweater and with.*

4. Incorrect Products: The model generates the incorrect products in the predicted utterance as compared to the one present in the original utterance as different products have similar attributes: Gold: *it can go well with unique branded, black colouring, chic type hand bag*; Predicted: *it can go well with black frame colour sunglasses.*

5. Wrong choice of images: The model focuses on incorrect images with respect to the conversational history due to the discussion over multiple images in history. Gold: *the upper material in the 2nd image is rubber lace*; Predicted: *the upper material in the 4th image is leather.*

7 Conclusion

In this paper, we have proposed an ordinal and attribute aware attention mechanism for natural language generation exploiting images and texts. In a multimodal setting, the information sharing between the modalities is significant for proper response generation, thereby leading to customer satisfaction. We incorporate the MFB fusing technique along with position and attribute aware attention mechanism for effective knowledge integration from the textual and visual modalities. On the recently released MMD dataset, the incorporation of our proposed techniques has shown improved performance for the task of textual response generation. In qualitative and quantitative analyses of the generated responses, we have observed contextually correct and informative responses, along with minor inaccuracies as discussed in the error analysis section. Overall the performance of our model shows the variations and more accurate responses in comparison to the other models keeping the attribute and position information of the generated responses intact.

In future, along with the opportunity of extending the architectural design and training methodologies to enhance the performance of our systems, we look forward to designing a specific component to enhance the natural language generation component of an end-to-end chatbot, by including image generation and retrieval systems for the completion of a multimodal dialogue system.

Acknowledgement

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018a. Improving context modelling in multimodal dialogue generation. *arXiv preprint arXiv:1810.11955*.

Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018b. A knowledge-grounded multimodal search-based conversational agent. *arXiv preprint arXiv:1810.11954*.

Peter Anderson, Stephen Gould, and Mark Johnson. 2018. Partially-supervised image captioning. In *Advances in Neural Information Processing Systems*, pages 1879–1890.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2612–2620.

Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, Fuming Ma, and Qi Ju. 2018. Improving image captioning with conditional generative adversarial nets. *arXiv preprint arXiv:1805.07112*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2951–2960.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on Artificial Intelligence and Statistics*, pages 249–256.

Shaojie Jiang and Maarten de Rijke. 2018. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. *arXiv preprint arXiv:1809.01941*.

Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Hung Le, S Hoi, Doyen Sahoo, and N Chen. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 801–809. ACM.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI2019 workshop*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics.

Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Thirty-Second AAAI Conference on Artificial Intelligence,pages 696-704*.

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*, 7(8).

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3776–3783. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1577–1586.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5456–5463. AAAI.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 196-205, Denver, Colorado. Association for Computational Linguistics*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *In Proceedings of ICML Deep Learning Workshop*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pages 2397–2406.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1821–1830.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.