RESEARCH ARTICLE

More to diverse: Generating diversified responses in a task oriented multimodal dialog system

Mauajama Firdaus^{1*}, Arunav Pratap Shandeelya², Asif Ekbal¹

¹ Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihar, India,
² Department of Electrical Engineering, International Institute of Information Technology Bhubaneswar, Orissa, India

* mauajama.pcs16,asif@iitp.ac.in

Abstract

Multimodal dialogue system, due to its many-fold applications, has gained much attention to the researchers and developers in recent times. With the release of large-scale multimodal dialog dataset Saha et al. 2018 on the fashion domain, it has been possible to investigate the dialogue systems having both textual and visual modalities. Response generation is an essential aspect of every dialogue system, and making the responses diverse is an important problem. For any goal-oriented conversational agent, the system's responses must be informative, diverse and polite, that may lead to better user experiences. In this paper, we propose an end-to-end neural framework for generating varied responses in a multimodal dialogue setup capturing information from both the text and image. Multimodal encoder with co-attention between the text and image is used for focusing on the different modalities to obtain better contextual information. For effective information sharing across the modalities, we combine the information of text and images using the BLOCK fusion technique that helps in learning an improved multimodal representation. We employ stochastic beam search with Gumble Top K-tricks to achieve diversified responses while preserving the content and politeness in the responses. Experimental results show that our proposed approach performs significantly better compared to the existing and baseline methods in terms of distinct metrics, and thereby generates more diverse responses that are informative, interesting and polite without any loss of information. Empirical evaluation also reveals that images, while used along with the text, improve the efficiency of the model in generating diversified responses.

OPEN ACCESS

Citation: Firdaus M, Pratap Shandeelya A, Ekbal A (2020) More to diverse: Generating diversified responses in a task oriented multimodal dialog system. PLoS ONE 15(11): e0241271. <https://doi.org/10.1371/journal.pone.0241271>

Editor: Haoran Xie, Lingnan University, HONG KONG

Received: February 29, 2020

Accepted: October 12, 2020

Published: November 5, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0241271>

Copyright: © 2020 Firdaus et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting information](#) files.

Funding: The research reported in the paper is partially supported by "I MP/2018/002072". All the

1 Introduction

Recent advancement in artificial intelligence (AI) has opened new frontiers in conversational agents. Human-machine interaction is an important application of AI that helps humans in their day-to-day lives. Progress in AI has led to the creation of personal assistants like Apple's Siri, Amazon's Alexa, Microsoft's Cortana that assist humans in their everyday works. The

fundings had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

capability of the machines to comprehend and complete users' goals has empowered researchers to build advanced dialogue systems. With the progress in visual question answering [1, 2] and image captioning [3, 4], the use of different modalities in dialogue agents has shown remarkable performance bringing the different areas of computer vision (CV) and natural language processing (NLP) together. Hence, multimodal dialogue system bridges the gap between vision and language, ensuring interdisciplinary research. Integration of information from different modalities, such as text, image, audio and video has known to provide complete information for building the effective end-to-end dialogue systems [5, 6].

Lately, several works on multimodal dialogue systems [7–9] have encouraged research in this direction by combining information from the different modalities, such as texts, audios, videos and images. Multimodal conversational systems provide completeness to the existing dialogue systems by providing necessary information that lacks in unimodal systems. The visual (in case of images and videos) and audio information help in building robust systems. Dialogue systems are grouped into two broad categories, namely open-domain conversational agents and goal-oriented dialogue systems. Response generation or Natural Language Generation (NLG), which handles the task of presenting the information to the user, is an important aspect of these systems. One of the Conversational AI System's main objectives is to combine language and vision for the development of robust dialogue systems.

1.1 Problem definition

The ability to present the information to the user is an important task of every dialogue system. Dialog manager decides on what to say to the user but how to say the information is the sole responsibility of natural language generation (NLG) module of a dialogue system. In our current work, we focus on response generation in a multimodal setup. In Fig 1, it is evident that the responses are dependent on the visual features as well as the textual information. Therefore, our present work is based on a recently released multimodal dialogue [5] dataset. Our primary focus is to generate diverse and polite responses in a multimodal system by utilizing the information of text and image. The existing systems concentrate on generating the responses, while in this work our target is to make the responses interesting and interactive to increase the customer satisfaction leading to customer retention. The task is challenging as we need to consider the information present in both the images and text, and to make the responses diverse and polite simultaneously. Existing unimodal systems suffer from the problem of generating dull and generic responses like Yes, No, Okay, I don't know etc. The researchers,

Fig 1. An example from the MMD dataset.

<https://doi.org/10.1371/journal.pone.0241271.g001>

therefore, focused on making the responses diverse [10, 11]. Our current work is greatly motivated from these prior works, but our focus is on building a multimodal dialogue system.

The user utterances play a significant role in building the dialogue context for generating coherent and appropriate responses, in accordance to the user demands. We focus on generating the textual responses only in a similar manner as [8, 12, 13]. Here, the task of multi-modal dialog generation is defined as follows: we consider both the modalities, i.e. text and images by effectively combining these two information for the generation of the textual response as opposed to the unimodal systems that typically use only textual information for generating the next response. The primary objective of this work is to generate the diversified and polite responses for increasing the human-machine interaction, and we do not focus on image retrieval or generation in our current work.

1.2 Motivation and contribution

Task-oriented dialogue systems are based primarily on textual (unimodal) information. Growing requirements in various fields, such as travel, entertainment, retail etc. require conversational agents to be able to communicate by incorporating information from the different modalities in order to build a robust system. Our research is based on the previously proposed multimodal dialog [5] (MMD) dataset, composed of the conversations related to e-commerce (fashion domain). The work is focused on generating textual responses based on both text and images present in the conversational history. An example from the MMD dataset is shown in Fig 1. It is evident from the figure that the responses are generic and not very interactive. It can be seen that the system's responses are repetitive and mostly in one word like yes or no. Hence, in our current research, we aim at generating the diverse textual responses by capturing information from both text and images. Diversification in textual responses is an open and exciting research problem [10, 14]. This becomes more challenging in multimodal dialogue systems, as information from the images is essential for making the responses diverse while preserving the content as there is a strong correlation between the text and images. For every goal-oriented dialogue systems, the interaction between the user and the system should lead to a better user experience and satisfaction. Hence, the generation of interactive, informative and polite responses is imperative for every conversational agent. Our primary objective is the generation of responses in accordance to the contextual information while making it more diverse and courteous.

In this work, we employ a multimodal hierarchical encoder to encode the information from both text and image. Due to strong dependency between the textual contents and images, we apply parallel co-attention in order to capture the effective evidences from image and text, and to create an useful context for the decoder. We employ BLOCK [15] fusion technique which is based on the block-superdiagonal tensor decomposition for learning better multimodal representation. For increasing the diversity in responses, we apply stochastic beam-search [16] which uses Gumbel Top k-trick for generating responses. To the best of our knowledge, we are the first in presenting a novel approach for making textual responses more diverse in a goal-oriented multimodal dialogue system.

The key characteristics of our present work are as follows:

- We employ a parallel co-attention mechanism to derive the dependencies between the text and image by focusing on the important textual and image information.
- We incorporate BLOCK—Bilinear superdiagonal fusion module to obtain the improved multimodal representations.

- We integrate Stochastic beam search with Gumbel Top k-trick for incorporating diversity in responses while preserving the contextual information.
- We achieve the state-of-the-art performance in making diverse and informative textual responses on the MMD dataset.

2 Related work

Response generation is a classical problem in every dialogue systems. Previously, unimodal systems having only text focused on generating responses that are informative, interesting and diverse. With the growth in Artificial Intelligence (AI) systems having images, audio and video modalities have been incorporated in making the robust dialogue systems. Below we present a brief overview of some of the works carried out in unimodal conversational agents followed by the multimodal systems.

2.1 Unimodal dialogue systems

Deep learning's efficacy has shown notable improvement in the generation of dialogue. As shown in [17, 18], deep neural models are quite successful in modelling the conversations. To capture the context of previous queries by the user, the authors in [19] proposed a hierarchical framework capable of preserving the past information. Similarly, to preserve the dependencies among the utterances, a hierarchical encoder-decoder framework was investigated in [20, 21]. Sequence to sequence (seq2seq) neural models often generate incomplete and boring responses, such as I don't know, Okay, Yes, No, etc. Hence, bringing the diversity in responses is an extremely challenging and interesting research problem for every conversational agent.

Generation of interesting responses has been the objective in many dialogue generation works, such as [10, 22–27]. In [10], maximum mutual information (MMI) as objective function was proposed for generating the diverse responses in neural models. Similarly, the authors in [23] used inverse token frequency as an objective function for generating the interesting responses. In [24, 25] and [27], conditional variational auto-encoders were used to generate coherent and diversified responses. In [22], adversarial learning was employed for generating informative and diverse responses. Deep reinforcement learning models [26] have also shown remarkable improvement in generating interesting responses. In [28], the inter-sibling ranking penalty was added to favour responses from diverse parents to be generated instead of using the standard beam search algorithm. The authors in [14] proposed diverse beam search algorithm that decodes a list of diverse outputs by optimizing for a diversity-augmented objective.

In [29], a dialogue generation model was proposed that directly captures the variability in possible responses to a given input, which reduces the 'boring/monotonous output' issue of deterministic dialogue models. The generative adversarial network was employed in [30] for generating informative and diversified text. In [31], SpaceFusion model was proposed to jointly optimize diversity and relevance of a sequence-to-sequence model with that of an auto-encoder model by leveraging regularization terms. In [32], the authors proposed a reinforcement learning-based approach which considers a set of responses jointly and generates multiple diverse responses simultaneously. The authors in [11] propose a Frequency-Aware Cross-Entropy (FACE) loss function for generating diverse responses by incorporating a weighting mechanism conditioned on token frequency. In [33], the authors proposed an easy-to-extend learning framework named MEMD (Multi-Encoder to Multi-Decoder), in which an auxiliary encoder and an auxiliary decoder are introduced to provide essential training guidance for generating diverse responses. A multi-mapping mechanism was proposed in [34] to

capture the one-to-many relationship, where multiple mapping modules are employed as latent mechanisms to model the semantic mappings from an input post to its diverse responses.

2.2 Multimodal dialogue systems

Recently, research in dialogue systems has shifted towards incorporating different modalities such as images, audio and video for capturing information to make the robust systems. The research reported in [7, 35–38] has been effective in narrowing the gap between vision and language. In [36], an Image Grounded Conversations (IGC) task was proposed, in which natural-sounding conversations were generated about a shared image. Similarly, the authors in [7] introduced the task of Visual Dialog, which required an AI agent to hold a meaningful dialogue with humans in natural, conversational language about visual content. In [39], a combination of generative adversarial networks (GANs) and reinforcement learning was employed to generate more human-like responses to the questions having visual information as well. The authors in [40] addressed the task of cross-modal semantic correlation for Visual dialog that utilized a dual visual attention mechanism for answering the questions. Similarly, for better visual and textual correlation, in [41], a textual-visual Reference-Aware Attention Network was employed for the generation of correct answers in accordance to the input image and dialog history. Our work varies from these as the Multimodal Dialog (MMD) conversations [5] deal with various images, and the growth in conversation depends on both image and text as opposed to a single image conversation.

Recently, video and textual modalities were investigated with the release of the DSTC7 dataset in [9] that made use of a multimodal transformer network to encode videos and incorporate information from the different modalities. Similarly in [6, 42, 43], DSTC7 dataset has been used for generation by incorporating audio and visual features. Earlier works on the MMD dataset reported in [12, 13, 44] used the hierarchical encoder-decoder model to generate responses by capturing information from text, images and the knowledge base. Recently, [8] proposed attribute and position-aware attention for generating textual responses. The authors in [45] used an hierarchical attention mechanism for generating responses on the MMD dataset.

Our work differs from the existing works (mainly, on the MMD dataset) in a sense that we aim here to generate diversified responses using stochastic beam search with Gumble k-tricks at the time of generation for a multimodal dialogue system. The task is challenging as we have to consider the images also for generating the coherent and informative responses. Hence, we focus on capturing vital contextual information by applying co-attention between text and image to achieve important details from both the modalities. We employ BLOCK fusion technique instead of the linear concatenation of modalities to obtain better multimodal representation. The end goal is to achieve responses that are not only coherent to the conversational history but also interesting, diverse and polite.

3 Methodology

In this section, we discuss the problem statement followed by the baseline and the proposed methodology.

3.1 Formal problem definition

In this paper, we address the task of generating diverse and polite textual responses in accordance to the conversational history in a multimodal dialog setting. The dialogues consist of textual utterance along with the multiple images. Given a context history of p turns, we address

Fig 2. Architectural diagram of our baseline multimodal hierarchical encoder-decoder.

<https://doi.org/10.1371/journal.pone.0241271.g002>

the task of generating the next response that is coherent, diverse and polite, leading to better and more engaging human-machine conversation. More precisely, given an user utterance $U_p = u_{p,1}, u_{p,2}, \dots, u_{p,j}$, a set of images $I_p = \text{img}_{p,1}, \text{img}_{p,2}, \dots, \text{img}_{p,j'}$, with the dialog history $H_p = (U_1, I_1), (U_2, I_2), \dots, (U_{p-1}, I_{p-1})$, we focus on generating interesting, informative, polite and context-aware response $Y_p = (y_{p,1}, y_{p,2}, \dots, y_{p,k})$ instead of template like generic and monotonous responses, such as *I don't know*, *Yes*, *No*, *Similar to...*, etc. This will enhance human-machine conversations by keeping the users engaged in the conversation. Here, p is the p^{th} turn of a given dialogue, while j is the number of words in a given textual utterance and j' is the number of images in a given utterance. Note that in every turn, the number of images $j' \leq 5$, so in-case of only text, vectors of zeros are considered in place of image representation.

3.2 Multimodal hierarchical encoder decoder

We construct a generative model for response generation, as shown in Fig 2, which is an extension of recently introduced Hierarchical Encoder-Decoder (HRED) architecture [20, 21]. As opposed to a standard sequence-to-sequence models [46], the dialogue context among the utterances is captured by adding utterance-level RNN (Recurrent Neural Network) over the word-level RNN increasing the efficacy of the encoder to capture the hierarchy in dialogue. The multimodal HRED (MHRED) is built upon the HRED to include text and image information in a single framework. The key components of MHRED are utterance encoder, image encoder, context encoder and decoder.

3.2.1 Utterance encoder. Given an utterance U_p , we use bidirectional Gated Recurrent Units (BiGRU) [47] to encode each word $n_{p,i}$ where $i \in (1, 2, 3, \dots, k)$ having d -dimensional embedding vectors into the hidden representation $h_{U,p,i}$ as follows:

$$\overrightarrow{h_{U,p,i}} = \text{GRU}_{U,f}(n_{p,i}, \overrightarrow{h_{U,p,i-1}}) \quad (1)$$

$$\overleftarrow{h}_{U,p,i} = GRU_{U,b}(n_{p,i}, \overleftarrow{h}_{U,p,i-1}) \quad (2)$$

$$h_{U,p,k}^{txt} = [\overrightarrow{h}_{U,p,i}, \overleftarrow{h}_{U,p,i}] \quad (3)$$

3.2.2 Image encoder. A pre-trained VGG-16 [48] having a 16-layer deep convolutional neural network (CNN) trained on more than one million of images present in the ImageNet dataset is used for encoding the images. It can classify images into 1000 object categories, such as dress, shoes, animals, keyboard, mouse, etc. As a result, the network can learn rich features from a wide range of images. Here, it is also used to extract the local image representation for all the images in the dialogue turns and concatenate them together. The concatenated image vector is passed through the linear layer to form the global image context representation as given below:

$$T_{p,i} = VGG(I_{p,i}) \quad (4)$$

$$T_p = Concat(T_{p,1}, T_{p,2}, \dots, T_{p,j'}) \quad (5)$$

$$h_{I,p}^{img} = ReLU(W_I T_p + b_I) \quad (6)$$

where W_I and b_I are the trainable weight matrix and biases. In every turn, the number of images $i \leq 5$, so in-case of only text, vectors of zeros are considered in place of image representation.

3.2.3 Context encoder. As shown in Fig 2, the final hidden representations from both image and text encoders are concatenated for each turn, and are given as input to the context level GRU. A hierarchical encoder is built to model the conversational history on top of the image and text encoder. The decoder GRU is initialised by the final hidden state of the context encoder.

$$h_{w,p}^{ctx} = GRU_w([h_{U,p,k}^{txt}; h_{I,p}^{img}], h_{w,p-1}) \quad (7)$$

where $h_{w,p}^{ctx}$ is the final hidden representation of the context for a given turn.

3.2.4 Decoder. In the decoding section, we build another GRU for generating the words sequentially based on the hidden state of the context GRU and the previously decoded words. We use input feeding decoding along with the attention [49] mechanism for enhancing the performance of the model. Using the decoder state $h_{d,t}^{dec}$ as the query vector, the attention layer is applied to the hidden state of the context encoder. The context vector and the decoder state are concatenated and used to calculate a final distribution of probability over the output

tokens.

$$h_{q,t}^{dec} = GRU_d(y_{p,t-1}, h_{q,t-1}) \quad (8)$$

$$\alpha_{t,m} = \text{softmax}(h_{w,p}^{ctx}{}^T W_f h_{q,t}) \quad (9)$$

$$c_t = \sum_{m=1}^k \alpha_{t,m} h_{w,p}^{ctx}, \quad (10)$$

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

$$P(y_t/y_{<t}) = \text{softmax}(W_s \tilde{h}_t) \quad (12)$$

where, W_f , W_s and $W_{\tilde{h}}$ are trainable weight matrices.

3.3 Proposed approach

To improve the performance of the MHRED model, we use an attention layer to mask out the insignificant information instead of merely concatenating the representations of the text and image encoder. It is essential to focus on both image and text modalities to achieve informative and coherent responses while generating exciting and diverse responses. For better representation across the modalities, we employ the recently proposed BLOCK fusion technique [15]. We provide the knowledge base information to the decoder for generating informative and context-aware responses. Finally, to generate diverse responses, we incorporate stochastic beam search at the time of generation. Beam search makes use of Gumbel max trick for top-K samples. The architectural diagram of the proposed framework is given in Fig 3.

3.3.1 Parallel Co-Attention (PCA). As opposed to visual question answering with single image in general, in multi-modal dialogue system we have multiple images over a context of turns. The MMD dataset has high correlation and dependency between the images and text. To achieve the correct attribute information in accordance to both text and image, it is essential to focus on both the modalities simultaneously. Hence in this work, to generate the context for every turn, we use the parallel co-attention as proposed in [50] to attend over the utterance encoder and image encoder simultaneously. In our case, we apply parallel co-attention between the textual utterance and the multiple images present in the utterance to attain significant attributes for the generation of more informative textual response. In the proposed framework, we connect the images and the utterance by computing the similarity between the utterance and image features at every pair of utterance-locations and image-locations. Since, parallel co-attention draws the essential features among both the text and visual counterparts hence in our work we employ PCA to obtain the important information from the text and the different images present in the input utterance. This facilitates the network in generating attribute centered responses as the attention focuses on the different attributes, such as color, pattern, shape, etc. of the products. The attention network, in accordance to text, captures the appropriate image information from multiple images. For example, the attended image is in consonance to the text such as 3rd image, 2nd image, etc. By achieving the relevant image information the generated response becomes more coherent to the dialogue history. The use of parallel co-attention in dialogue systems is new, especially for multi-modal dialogue setting and under a situation when multiple images exist in a single turn of a dialogue utterance. Precisely, given an utterance representation $U \in \mathbb{R}^{d \times K}$, image representation $I \in \mathbb{R}^{d \times N}$, the affinity

Fig 3. Architectural diagram of our proposed multi-modal hierarchical encoder-decoder with parallel co-attention mechanism along with BLOCK fusion technique between the text and image. Input to the decoder is the attended and fused context vector along with the decoder and KB inputs.

<https://doi.org/10.1371/journal.pone.0241271.g003>

matrix $C = \mathbb{R}^{K \times N}$ can be computed as:

$$C = \tanh(U^T, W_b, I) \quad (13)$$

where W_b are the trainable weights. After the computation of the affinity matrix C , one of the plausible ways of calculating the utterance or image attention distribution is to simply maximize the affinity over the locations of other modality, i.e. $\alpha^U[n] = \max_i(C_{i,n})$ and $\alpha^I[n] = \max_j(C_{j,n})$. In contrast to choosing the max activation, the performance can be further improved by considering the affinity matrix C as a feature. Therefore, the attention feature map for parallel co-attention between image and text utterance can be represented as follows:

$$h_I = \tanh(W_I I + (W_U U)C) \quad (14)$$

$$h_U = \tanh(W_U U + (W_I I)C^T) \quad (15)$$

$$\alpha^I = \text{softmax}(W_{h,I} h_I) \quad (16)$$

$$\alpha^U = \text{softmax}(W_{h,U} h_U) \quad (17)$$

where $W_{h,b}$, $W_{h,U} \in \mathbb{R}^S$ are the weight parameters. Here, $\alpha^I \in \mathbb{R}^N$ and $\alpha^U \in \mathbb{R}^K$ are the attention probabilities of each image and word, respectively. The affinity matrix C transforms utterance attention space to image attention space (vice versa for C^T). Based on the above attention

weights, the utterance and image vectors are calculated as the weighted sum of the utterance and image features given by:

$$U_a = \sum_{n=1}^N \alpha_n^U U_n \quad (18)$$

$$I_a = \sum_{t=1}^T \alpha_t^I I_t \quad (19)$$

We employ parallel co-attention between the text utterances and images to obtain a more focused information conditioned on the text utterances. We concatenate U_a and I_a , where U_a is the final utterance representation and I_a is the final image representation.

3.3.2 BLOCK fusion technique. Multi-modal representation learning is an important aspect of every multi-modal system. One of the very fundamental principles is to fuse the information from the different modalities (e.g. text, audio and visual) effectively for developing a robust dialogue system. In our case, we need better representation to learn the relationships between the textual and visual modalities. Previous works on multi-modal dialogue systems [12, 13, 44, 45] have merely concatenated the textual and visual representation, and thus have only linear interaction between the modalities. This is insufficient in capturing the complex interactions between the diverse modalities. For better non-linear interaction between the two modalities, we apply BLOCK fusion technique. With recent success of BLOCK fusion technique in visual question answering [15], we tend to incorporate it in case of multi-modal dialogue system, where the interaction between the modalities is highly intricate as the previous dialogue history along with multiple images needs to be considered for generating appropriate responses that are contextually coherent with both textual and visual modalities. With this robust fusion technique, i.e. BLOCK [15], we tend to capture better multi-modal representation. This is based on the concept of block-superdiagonal tensor decomposition. It utilizes the concept of block-term ranks that generalize the notion of both rank and mode-ranks for tensors. It helps in defining new approaches for optimizing the trade-off between the complexity and expressiveness of the fusion model. Hence, it can capture and represent very intricate interactions between the modalities while handling powerful uni-modal representations.

The final image representation I_a and final utterance representation U_a serve as the inputs to the BLOCK module. The BLOCK fusion approach brings state-of-the-art performance for effective interaction between the two modalities, and hence enhances the performance of the system. Therefore, we employ it in our multi-modal encoder to obtain better representation through effective interaction between the textual and visual modalities. In our case, the BLOCK module is computed as:

$$\zeta = C((U_a^T W_U) * (I_a^T W_I)) \quad (20)$$

where W_U and W_I are the trainable parameters, and $*$ denotes the element-wise product.

Hence, the final multi-modal BLOCK fusion between the final textual representation, final image representation and contextual representation is computed as:

$$h_d = \text{BLOCK}(U_a, I_a) \quad (21)$$

3.3.3 Knowledge base (KB). The knowledge base encoder used in our framework is the same as [13]. The knowledge base of the MMD dataset contains information about the contextual queries and celebrities endorsing various products and brands. Hence, to provide this

additional information to our proposed model, we employ self-attention on KB input to achieve more focused information as follows:

$$h_s^{query} = n_p^{query}(h_{n-1}^{query}, k_{l,t}) \quad (22)$$

$$h_f^{ent} = n_p^{ent}(h_{n-1}^{ent}, d_{l,t}) \quad (23)$$

$$h_{net}^{kb} = [h_s^{query}, h_f^{ent}] \quad (24)$$

$$\beta = \text{softmax}(W_{x1} \tanh(W_{v2} h_{net}^{kb})) \quad (25)$$

where W_{x1} and W_{v2} are trainable parameters.

We use the attended KB output along with the decoder input as the combined input at each time step of the decoder.

As the knowledge base (KB) input remains intact for a particular dialogue context, therefore we concatenate the KB input with the decoder input in a similar manner as in [13]. Apart from that, we fuse the text and visual representation at the encoder level to form the context for every dialogue turn. During decoding, to make the responses diverse, we use stochastic beam search [16] for diversification and informative responses in accordance to the contextual information.

$$h_{d,t}^{dec} = f^{dec}(h_{d,t-1}^{dec}, W_{d,t}, h_{w,p}^{ctx}, h_{net}^{kb}) \quad (26)$$

3.4 Training and inference

The whole model is trained using teacher-forced cross entropy [51] at every decoding step to minimize the negative log likelihood on the model distribution. We define $\hat{y} = \hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_m$ as the ground truth of the given input sequence.

$$J(\theta) = -\sum_{s=1}^n \log p(\hat{y}_s | y_1, y_2, \dots, y_{m-1}) \quad (27)$$

For diversity, we use Stochastic beam search for generating the response, which is more diversified and also capable of preserving the contextual information. Stochastic beam search was derived from the Gumbel Top-K trick to sample sequences without replacement. This requires instantiating all the sequences in the domain to get the largest perturbed probability. Then we transit to a top-down sampling of the perturbed log probabilities.

3.4.1 Gumbel top K tricks. As mentioned in [16], the model is represented as a tree where internal nodes at level p represent the partial sequences $y_{1:p}$ and the nodes corresponding to leaf nodes are the completed sequences y^i . We identify the leaf nodes by its index $i \in (1, 2, \dots, n)$ and write y^i as the corresponding sequence with log probability $\phi_i = \log p_{\theta}(y^i)$. The Gumbel max-trick samples from this distribution by independently perturbing the log probabilities with Gumbel noise and finding the largest element. Now, this generalizes the Gumbel Max Trick to Gumbel Top-K tricks to sample with the size K without any replacement, by taking the indices of k-largest perturbed log-probabilities.

3.5 Baseline models

For our experiment, we implement the following baseline models.

Model 1 (MHRED): The first model is the baseline MHRED model as previously described in the methodology section. This is shown in Fig 2, that has utterance encoder, image encoder, context encoder and the decoder for generating the responses.

Model 2 (MHRED+KB): In this model, we add the KB information at the decoder to incorporate contextual queries and celebrity information for endorsing different brands and products for informative response generation that is in accordance with the conversational history.

Model 3 (MHRED+KB+A): In this model, we employ global attention [49] at the decoder for better decoding at each time step.

Model 4 (MHRED+KB+A+DAA(I,T)): In this model, for different attributes on the text, we implement dynamic attribute attention to obtain the enhanced representation computed by:

$$\beta_t = \text{softmax}(W_p^T H_u) \quad (28)$$

$$U_p = \beta_t H_u^T \quad (29)$$

Here, a self-attended text embedding is used as query U_p to calculate the attention distribution over the image feature representation as $H_I = [H_{I,1}, H_{I,2}, H_{I,3}, \dots, H_{I,j}]$.

$$\beta_k = \text{softmax}(U_p^T W_{p'} H_I) \quad (30)$$

$$I_a = \beta_k H_I^T \quad (31)$$

Model 5 (MHRED+KB+A+PCA(I,T)): In this model, we employ parallel co-attention mechanism as described in 3.3.1 to attend the image and text simultaneously. We connect image and text by calculating the similarity between the image and text features at all the pairs of image and text utterances, as previously described in the methodology section.

Model 6 (MHRED+KB+A+PCA(I,T)+MFB(I,T)): In this model, we concatenate pairwise text and image features U_f and I_f obtained after parallel co-attention mechanism as input to the MFB module for better interaction between the modalities as used in [8].

Model 7 (MHRED+KB+A+PCA(I,T)+BLOCK(I,T)): In this model, we concatenate the pairwise output of text and image features U_a and I_a obtained after parallel co-attention mechanism as input to the BLOCK module where the output of BLOCK serves as input to the context encoder.

4 Dataset

Our research is based on the Multi-modal Dialog (MMD) dataset [5] consisting of 150k chat sessions between the customer and sales representative. During the sequence of customer-agent interactions, domain-specific information in the fashion domain was collected. The dialogues easily integrate text and image knowledge into a conversation that brings together different modalities to create a sophisticated dialog system. The dataset presents new challenges for multi-modal, goal-oriented dialogue systems having complex user sentences. The detailed information of the MMD dataset is presented in Table 1. The authors [5], for experimentation unroll the different images to incorporate only one image for a single utterance. This method, though computationally learns, eventually lacks the goal of capturing multi-modality over the context of multiple images and text. Therefore, in our study, we use a different version of the dataset as outlined in [12, 13] to capture a large number of images as the concatenated context vector for each turn of a dialogue. The motivation behind this is the fact that multiple images are required for providing the correct responses to the users. As shown in the example before,

Table 1. Dataset statistics of multi-modal dialogue (MMD) dataset.

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

<https://doi.org/10.1371/journal.pone.0241271.t001>

can you show me something similar like the 2nd image?, the different images present play an important role for the generation of contextually correct responses.

For incorporating diversity and politeness in the generated responses, we manually annotated 40% of the training set dialogues with courteous phrases and varied responses that are contextually coherent with the dialogue history and the product being discussed in the conversation. Three annotators proficient in English language were assigned to annotate the textual responses making it more empathetic and diverse by incorporating phrases for apology, appreciation, assurance and greetings. We observe the multi-rater Kappa agreement ratio of approximately 80%, which may be considered as reliable. The annotated courteous data was used for training the model for making it more polite and diverse. An example of the polite version of the MMD dataset annotated by experts is shown in Fig 4.

5 Experiments

In this section, we present the information about the implementation details of our proposed framework.

5.1 Implementation details

All the implementations have been performed using PyTorch <https://pytorch.org/> framework. We use 512-dimensional word embedding initialized randomly. We did not use any delexicalisation, and our model learns independent of the context encoder and knowledge base (KB). All the encoder and decoder have 1-layer GRU cell with 512 hidden dimensions. For image representation, we use 4096-dimensional FC6 layer from VGG-16 [48], pre-trained on the Imagenet. We use AMSGrad [52] as the model training optimizer to alleviate the problems of slow convergence. We use dropout [53] with the probability of 0.4 to avoid over-fitting. All the parameters are randomly initialized using Gaussian distribution with Xavier scheme [54]. For generating diversified responses, we use the different methods, such as beam search, diverse beam search and greedy sampling. Finally, in our proposed model, we employ stochastic beam search with sample size k: 5, 10 as implemented in fairseq <https://ai.facebook.com/tools/fairseq/>. It works based on the Gumbel Top-k trick that samples sequences without replacement from the sequence model. We experiment with different learning rates and lastly fix it as 0.0004.

To measure the politeness quotient in the generated response we design a politeness classifier as presented in Fig 5. The input is first converted into embeddings using the embedding layer. We use 300-dimensional Glove embedding for representing the utterances. The embedded utterance is fed to the convolutional layer with filter size 3. After convolution, we apply max pooling to obtain the hidden representations. For more abstract representation of the dialogue utterance, we apply a unidirectional LSTM network on the hidden representation followed by self-attention. The number of neurons on the LSTM layer is fixed to be 200. Finally, we apply the softmax layer for attaining the politeness accuracy of the given utterance. To

U1: Hi there!
 S1: Hello! how I can help you?
 U1: Can you show me some trousers for my father-in-law.
 S1: The similar one looking are

U1: What is the colour and material in the 4th result?
 S1: It has grey colour cotton material.
 U1: Is the 3rd image good for dry cleaning?
 S1: No.

U1: I would like to see a few waist-length Hoodies for myself.
 S1: Showing you in just a minute.

U1: Will gloves complement any of these?
 S1: Yes
 U1: Do you have the feeling that scarf with fabric type prints will go well with any of these?
 S1: Yes
 U1: What about the full lengthed leggings?
 S1: No.

Orininal MMD dataset

U1: Hi there!
 S1: Hello please tell me, anything that I can help you with?
 U1: Can you show me some trousers for my father-in-law.
 S1: Sure, thanks for sharing the information, here are some images for your reference

U1: What is the colour and material in the 4th result?
 S1: The trackpants in the 4th image has grey colour cotton material.
 U1: Is the 3rd image good for dry cleaning.
 S1: Sorry, but we won't suggest the 3rd image for dry clean.

U1: I would like to see a few waist-length Hoodies for myself.
 S1: Please have patience, we are showing you in just a minute.

U1: Will gloves complement any of these?
 S1: Yes they will suit the hoodies really well.
 U1: Do you have the feeling that scarf with fabric type prints will go well with any of these?
 S1: Yes printed scarfs will be well suited with these.
 U1: What about the full lengthed leggings?
 S1: I am sorry but I don't think it will be good with the hoodies.

Polite and Diversified MMD dataset

Fig 4. An example of the new diversified and polite version of the MMD dataset.

<https://doi.org/10.1371/journal.pone.0241271.g004>

avoid over-fitting of the neural network a useful regularisation technique known as dropout has been used in our model. At the time of forward propagation, the neurons are randomly tuned-off so that the convergence of weights is restricted to the identical positions. For optimisation and regularisation, we use Adam optimiser along with 15% dropout in our model. Categorical cross-entropy is employed to update the model parameters.

6 Results and discussion

In this section, we present the evaluation metrics (automatic and human), report the experimental results along with the detailed analysis and comparisons.

6.1 Evaluation metrics

6.1.1 Automatic evaluation metrics. To evaluate the model at relevance and grammatical level, we report the results using the standard metrics like:

Fig 5. Architectural diagram of politeness classifier.

<https://doi.org/10.1371/journal.pone.0241271.g005>

- **Perplexity [55]:** Perplexity has been used to test our model at the content level. Smaller scores of perplexity mean the responses generated are more grammatical and fluent.
- **BLEU-4 [56]:** BLEU measures the n-grams overlap between the generated response and the gold response, and has become a common measure for comparing task-oriented dialog systems. It is used to measure the content preservation in the generated responses. BLEU is weakly correlated with human judgments [57], but a low BLEU score across different models suggests that a dataset is highly complex.
- **Distinct-1 and Distinct-2:** We report the degree of diversity by calculating the number of distinct unigrams (distinct-1) and bigrams (distinct-2) in the generated responses. The resulting metric is, thus, a type-token ratio for unigrams and bigrams as mentioned in [10]. This metric measures the number of distinct k-grams in the replies generated and is scaled by the total number of tokens generated to avoid long replies. This metric is an indicator of word-level diversity for the generated responses.

- **Politeness Accuracy:** We also compute politeness score using a pre-trained classifier as shown in Fig 5 (trained on Stanford Politeness corpus) for measuring the degree of politeness in the generated responses similar to [58]. The classifier takes as input the generated response and predicts a probability value giving us the politeness accuracy of the generated response.

6.1.2 Human evaluation metrics. We adopt human evaluation metrics to compare the efficiency of distinct models in order to examine the quality of the generated responses. We randomly sample 1500 responses from the test set for human evaluation. Three human annotators with post-graduate exposure were assigned to evaluate the generated response in terms of the various human evaluation metrics as mentioned below. Given an utterance, images along with the conversation history, annotators were asked to evaluate the correctness, polite preservation, diversity and relevance of the responses generated by the different models.

- **Fluency (F):** The generated response is grammatically correct and is free of any errors.
- **Diversity (D):** It is used to judge whether the generated response is diverse in comparison to the ground truth.
- **Relevance (R):** The response generated is focused on the aspect being discussed (style, colour, material, etc.) and is in accordance with the conversational history.
- **Politeness (P):** The generated responses are polite and in accordance to the context having both text and images.

We follow the scoring scheme for fluency and relevance as- 0: incorrect or incomplete, 1: moderately correct, and 2: correct. The scoring scheme for diversity and politeness preservation is 0: for the absence of diversity and politeness in the reply and 1: for the presence of diversity and politeness in the response. We compute the Fleiss' kappa [59] for the above metrics to measure inter-rater consistency. The kappa score for fluency and relevance is 0.75, and for diversity and politeness 0.77, indicating substantial agreement.

6.2 Results of automatic evaluation

We present the results of our experiments in Table 2. It is evident from the table that our proposed approach outperforms the baselines in generating diversified responses, and these improvements are statistically significant. We perform statistical significance t-test [60] and it is conducted at 5% (0.05) significance level. The perplexity score of the proposed approach with $k = 5$ is the least, thereby proving the fact that the generated responses are better from the rest. This suggests that our proposed framework employing stochastic beam search with gumble top k-tricks (here $k = 5$) is able to generate better responses in contrast to the baseline models and the existing system. In the baseline approaches, the framework employing PCA and BLOCK fusion technique outperforms all the other baseline networks. This indicates that both PCA and BLOCK fusion employed together makes the network perform better as opposed to the frameworks in which they are employed individually. Similarly, there is a decrease in perplexity of about 15 points when we use BLOCK instead of MFB fusion technique that states the superiority of BLOCK fusion in comparison to the MFB fusion technique.

Similarly, the distinct-1 and distinct-2 metrics of the final model with stochastic beam search decoding proves that the proposed framework have been successful in generating diverse responses for the MMD dataset. Our proposed framework outperforms the existing approaches such as MMI [10], MMI-antiLM [10] and MMI-Bidi [10] for both distinct-1 and

Table 2. Results of different models on MMD dataset.

	Model Description	Perplexity (PPL)	BLEU-4	Distinct-1 (d-1)	Distinct-2 (d-2)	Politeness Accuracy (PA)
Existing Approaches	Agarwal et. al. [13], (MHRED + Attn)	1.098	0.4451	-	-	-
	Chauhan et. al. [8], (MHRED + Attn)	1.082	0.4454	-	-	-
	Li et. al. [10], (MMI)	1.789	-	0.0315	0.0865	-
	Li et. al. [10], (MMI—AntiLM)	1.980	-	0.0201	0.0906	-
	Li et. al. [10], (MMI-bidi)	1.967	-	0.0244	0.0954	-
Baseline Approaches	MHRED	1.009	0.4561	0.0384	0.1201	0.76
	MHRED+KB	1.009	0.4591	0.0435	0.1595	0.76
	MHRED+KB+A	1.008	0.4624	0.0405	0.1453	0.77
	MHRED+KB+A+DAA(I,T)	1.005	0.4715	0.0352	0.1205	0.78
	MHRED+KB+A+PCA(I,T)	1.004	0.4875	0.0429	0.1138	0.78
	MHRED+KB+A+PCA(I,T)+MFB(I,T)	1.0036	0.4965	0.0358	0.0981	0.78
	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)	1.0021	0.4981	0.0399	0.1089	0.80
Proposed Approaches	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 5)	1.0008	0.4691	0.0547	0.1653	0.84
	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 10)	1.0009	0.4359	0.0501	0.1393	0.81

<https://doi.org/10.1371/journal.pone.0241271.t002>

distinct-2 metrics. From the baseline models, there is an improvement in the distinct-1 and distinct-2 measures for the proposed approach having stochastic beam search decoder for both $k = 5$ and 10 . This implies the efficacy of the stochastic beam search that ensures diverseness in the responses. The distinct-1 and distinct-2 scores demonstrate a huge jump in comparison to the best performing baseline network $MHRED+KB+A+PCA(I,T)+BLOCK(I,T)$, thereby illustrating the effectiveness of stochastic beam search with Gumble k-tricks. From the table it is visible that our baseline framework shows improved performance in case of BLEU-4 metric in comparison to the existing systems. We observe that there is a drop in BLEU-4 metric in our proposed model, $MHRED+KB+A+PCA(I,T)+BLOCK(I,T)$ compared to the baseline. This favors our assumption that the generated response is diverse in nature, and hence is not completely similar to the ground-truth. The politeness score also increases in our proposed framework with an increment of 4% with respect to the final baseline model. This improvement indicates that the generated responses are more polite than the baselines. Therefore, our proposed approach ensures that in the multi-modal setup, both diversity and politeness are preserved.

6.3 Analysis of network parameters

We provide the analysis of different hyper-parameters such as dropout and learning rate for our proposed framework. We used Perplexity (PPL) as the primary metric to fine-tune the framework for determining the network parameters in case of dropout and learning rate. After fixing these parameters we do the complete evaluation in terms of both automatic and human metrics for all the baselines and proposed framework. We used politeness accuracy (PA) as the primary metric for determining the number of LSTM layers in the politeness classifier. The dropout probability was determined to be 0.4 by considering the range of 0.1—0.8. The statistical analysis of the results for the different values of dropout probability for the proposed framework is provided in Table 3 of the revised manuscript. In Fig 6(a) we depict the performance of the proposed framework in case of dropout. It is evident that the proposed framework performs best when the dropout probability is 0.4. Therefore, for all the experiments (all baselines)

Table 3. Analysis on proposed framework for learning rate and dropout.

Proposed Framework	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 5)							
Learning Rate	0.0010	0.0008	0.0006	0.0005	0.0004	0.0003	0.0002	0.0001
PPL	1.0036	1.0021	1.0019	1.0013	1.0008	1.0017	1.0022	1.0020
Dropout	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
PPL	1.0054	1.0031	1.0015	1.0008	1.0011	1.0013	1.0014	1.0012

<https://doi.org/10.1371/journal.pone.0241271.t003>

and metrics we fix the dropout probability as stated. For learning rate we determined the value to be 0.0004 similarly as [12, 13, 45]. Also, we cross-verified by taking a range of 0.001 to 0.0001 to determine the learning rate in case of our proposed framework. As shown in Table 3 and Fig 6(b), the proposed framework performs best when the learning rate is fixed at 0.0004 in a similar manner as the existing literature.

For the politeness classifier, we had fixed the LSTM layers at 200 by evaluating the performance of the classifier for the responses generated by our proposed framework. We checked the performance of the classifier by using different layers of LSTM as shown in Table 4 of the

Fig 6. Analysis of different network parameters on the proposed framework.

<https://doi.org/10.1371/journal.pone.0241271.g006>

Table 4. Analysis of LSTM layers on politeness classifier for the proposed framework.

Proposed Framework	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 5)					
LSTM Layers	50	100	150	200	250	300
Politeness Accuracy	0.79	0.80	0.82	0.84	0.82	0.82

<https://doi.org/10.1371/journal.pone.0241271.t004>

revised manuscript. From the Fig 6(c), it is evident that by increasing the number of LSTM layers the performance of the classifier improved with the maximum accuracy being 0.84 in case of 200 layers. On increasing the number of layers the performance declined and eventually became constant with accuracy lesser than the best performing classifier at 200 layers. This analysis helped in deciding the number of LSTM layers for the classifier.

6.4 Comparisons to the existing systems

In Table 2, we provide the results of our proposed framework in comparison to the existing state-of-the-art methods. For response generation in multi-modal systems, we compare our current work with the existing systems [8, 13]. For a fair comparison with the existing systems, the data used for experimentation should be similar in terms of its structure, genre as well as annotation in order to draw correct conclusions on the results obtained. Hence, we compare our proposed approach with these systems as the dataset is identical. Also, both the existing systems used for comparing with our current system focuses on the task of textual response generation in multi-modal dialogue systems. In [13], the authors employed a knowledge base for generating the textual responses while using both the textual and visual information. Similarly, in [8], the authors have exploited the position and attribute information in text and images for generating the textual responses. As it is evident from Table 2, our proposed system outperforms the existing systems significantly in terms of Perplexity and BLEU-4 metrics. By applying parallel co-attention and BLOCK fusion technique, our best performing baseline framework achieves an improvement of around 5% in BLEU score from both the existing approaches. This is mainly due to the fact that the parallel co-attention network simultaneously captures the essential information from the entire utterance possessing textual and visual knowledge. We also attain enhanced representation of the utterance by having improved interaction between the modalities with the help of BLOCK fusion technique. One of the purposes of the current research is to accomplish diverse response generation. Evaluation shows that by applying stochastic beam search, we observe an improvement of 2% in BLEU score (with $k = 5$) as opposed to [8] and [13]. The increase in BLEU score signifies that the responses are diverse and different from the ground-truth responses, thereby, helping in achieving the desired task. Also, the decrease in perplexity score testifies that the generated responses are grammatically correct and better in comparison to all the baselines and the existing frameworks. The other existing systems on multi-modal dialog systems focus on the tasks of response generation and image retrieval both, which does not come within the scope of our work. Moreover, the datasets used in these works are different and therefore correct conclusions cannot be drawn for fair comparisons.

It is to be noted that as this is the very first work that focuses on generating diverse responses in a multi-modal setup, hence we are unable to compare our work with any other systems with respect to the diversity. Since the primary goal of this research is to bring diversity in responses, therefore we compare our proposed system with the state-of-the-art techniques [10] for generating diverse responses. In the existing methods the authors employed various objective functions, such as maximum mutual information (MMI), anti language models (anti-LM), MMI-bidi approach for making the responses diverse. We analyze the

Table 5. Results of human evaluation on different models for the MMD dataset.

	Model Description	Fluency			Relevance			Diversity		Politeness Preservation	
		0	1	2	0	1	2	0	1	0	1
Baseline	MHRED	25.62	38.54	35.84	42.14	35.98	21.88	57.13	42.87	44.73	55.27
	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)	22.14	40.22	37.64	31.66	38.45	29.89	54.22	45.78	41.69	58.31
Proposed	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 5)	13.81	44.88	41.31	26.59	38.54	34.87	25.68	74.32	38.07	61.93
	MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K = 10)	14.35	45.17	40.48	28.37	37.32	34.31	27.84	72.16	39.77	60.23

<https://doi.org/10.1371/journal.pone.0241271.t005>

responses generated by these techniques on the MMD dataset to have a thorough comparison in terms of distinct-1 and distinct-2 metrics. From the table, it is visible that our proposed framework performs better than the existing approaches for the task of diverse response generation. In contrast to the existing system and the baseline approaches, our proposed framework is capable of generating responses that are diverse in nature, thereby increasing the interactive property of the responses.

6.5 Human evaluation results

In Table 5, we present the human evaluation results for the baseline as well as the proposed approaches. The fluency of the proposed model is better in comparison to both the baseline models. We observe a fluency score of 41.31 for the proposed approach with stochastic beam search decoder ($k = 5$) along with BLOCK fusion techniques and attention modules. Hence, the models are capable of generating fluent responses. Also, there is an improvement of 4.98 in the relevance score over the baseline model. In case of diversity, it is evident from the table that the stochastic beam search decoder with $k = 5$ has achieved a score of 74.32, thereby, proving the fact that the responses generated are interesting and diverse in comparison to the baselines. Finally, the politeness score of the proposed approach is also better than the baseline model. Hence, the human evaluation results conclude that the proposed method has been successful in generating fluent, relevant, diverse as well as polite responses in a multi-modal dialogue system. In Fig 7, we present the responses generated by the different models. As it is

Context	I would love to buy few moccasins having suede heels.

Input	Will exclusive styled t-shirt pair well with loafers with suede type heels?
Ground-truth	Yes
Baseline Response: (MHRED)	Yes
Baseline Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T))	Yes it will
Proposed Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K=5))	Nice choice, it will be a great pair.
Proposed Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K=10))	Yes sure it will be nice.

Example 1

Context	I am interested in looking at cufflinks for myself.
	Absolutely.
Input	Do you think the delicate style cufflinks in the 1st image go with tuxedo?
Ground-truth	No
Baseline Response: (MHRED)	Yes
Baseline Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T))	No
Proposed Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K=5))	I am sorry, but this will not look good together.
Proposed Response: (MHRED+KB+A+PCA(I,T)+BLOCK(I,T)+SBS(K=10))	No I don't think so.

Example 2

Fig 7. Examples of responses generated by different models along with contextual information and input.

<https://doi.org/10.1371/journal.pone.0241271.g007>

visible from the figure, the responses generated are quite diverse and completely fluent. Also, instead of dull and direct Yes and No responses, the model has learned to be more polite by appreciating and apologizing, whenever required, according to the conversational history.

6.6 Attention analysis

For the model to be capable of learning different attribute information, we incorporate the knowledge base (KB) information in our model in the similar way as was done in [13]. To obtain a more focused information from KB, we apply self-attention on the KB inputs. The attention visualization of the KB is given in Fig 8. From the figure, it is noticeable that the model can focus on the relevant attributes for better dialogue generation. In Example 1 from the figure, it is evident that eyeglasses of brand Vincent is given more focus as the brand is an important feature for the user. Similarly, in Example 2 the print of the shirt is a significant aspect for the selection of appropriate shirt for the user. From these examples, it is clear that the KB input facilitates in capturing the correct attribute of a product, thereby, enhancing the quality of the generated response by making it more coherent to the user needs. Different attributes, such as the *brand*, *print*, *type*, *name* of a product gets attended by the self-attention applied on the KB input which is in accordance to the dialogue context as observed by the attention visualization in Fig 8 assists the generation of the textual responses.

For better interaction between the image and text, we employ parallel co-attention in our proposed framework as described in Section 3.3.1 of the Methodology section. For a complete understanding of the parallel co-attention and its effectiveness we provide an attention visualization of the parallel co-attention module in Fig 9. The attention on the text and the corresponding image is needed for better response generation. By employing parallel co-attention it

Example 1

name:eyeglasses shape round brand Vincent
image_type rear

Example 2

name:shirt fits loose brand Blackcherry gender men
image_type front print plain colour blue

Example 3

name denim shorts name three-fourths name bermuda shorts look gender all
gender women

Example 4

name floater name floaters image_type front image_type look gender all gender
women synsets floater synsets floaters print liberty

Example 5

name watch name watches image_type front image_type look gender all gender
men material leather synsets watch synsets watches type party

Fig 8. Attention map for KB Input.

<https://doi.org/10.1371/journal.pone.0241271.g008>

Example 1:

U1: What is the colour and material in the 4th result?

S1: It has grey colour cotton material.

Example 2:

U1: what material is it in the 2nd image?

S1: The brogues in the 2nd image has leather and rubber material

Fig 9. Attention map for parallel co-attention between the text and image.

<https://doi.org/10.1371/journal.pone.0241271.g009>

can be observed from the attention visualization that the attention mechanism simultaneously focuses on the given text and the images, providing an enhanced representation to the decoder for the generation of better responses. From the figure, it is quite clear that the attributes and different images being discussed in text are also getting focused on the corresponding image. We thereby provide more information and better evidences to the decoder for generating the response. In the first example from the Figure, the 4th image gets attended as it is being discussed in the textual utterance as well. The color and material of the image and the information in the textual part of the utterance, such as cotton simultaneously gets attended providing improved representation of the entire utterance having both textual and visual features. Similarly, in Example 2 from the set of 5 images the 2nd image gets selected and the attributes such as leather, rubber material gets attended in the textual part of the utterance. From the figure showcasing the attention visualization, it can be concluded that the parallel co-attention mechanism assists in capturing the important and critical features of both the visual and textual part of the utterance simultaneously.

6.7 Error analysis

We closely analyze the outputs of the generated response so that we are aware of the errors made by the proposed dialogue generation framework. The common errors made by the model are:

- **Additional Information:** The model sometimes generates additional/extra information in case of attributes for a few products. For example, **Gold:** *The material of the trousers is cotton.*, **Predicted:** *The trousers have cotton polyester material with check patterns.*
- **Incorrect information:** The model is incapable of presenting the correct celebrity information in most of the cases. For example, **Gold:** *Celebrities cel_205, cel_2254 and cel_101 endorse this kind of shoes.*, **Predicted:** *This kind of shoes is endorsed by celebrity cel_123.*

- **Erroneous image selection:** The model is incompetent in selecting the images having contextual information of more than 5 turns, thereby generating incorrect responses in some cases. There are also the cases, where due to the discussion of multiple images in the conversational history, wrong images get selected, making the responses incorrect.

7 Conclusion and future work

In this paper, we have proposed a novel approach for diversifying textual responses in a multi-modal dialogue system. Our method makes use of parallel and dynamic attention to focus on image, text and knowledge base to capture the information present in both the modalities. BLOCK fusion technique was incorporated in the multi-modal encoder to obtain better representation from the modalities by improving the interaction between them. Experimental results prove that the attention mechanism and Block fusion technique help in generating correct and informative responses. For diversification, we have employed stochastic beam search with Gumble k-tricks. Detailed empirical analysis shows that our proposed model is not only capable of generating informative responses, but also the responses are diverse and polite.

In future, we would take an opportunity of extending the architectural design to enhance the performance of our system. Also, we would focus on image retrieval and generation for building an end-to-end framework for multi-modal dialogue systems.

Supporting information

S1 File.
(ZIP)

Author Contributions

Conceptualization: Asif Ekbal.

Data curation: Mauajama Firdaus, Arunav Pratap Shandeelya.

Funding acquisition: Asif Ekbal.

Investigation: Mauajama Firdaus, Arunav Pratap Shandeelya, Asif Ekbal.

Methodology: Mauajama Firdaus, Arunav Pratap Shandeelya.

Project administration: Asif Ekbal.

Software: Arunav Pratap Shandeelya.

Supervision: Asif Ekbal.

Validation: Arunav Pratap Shandeelya.

Writing – original draft: Mauajama Firdaus.

Writing – review & editing: Asif Ekbal.

References

1. Wu Q, Teney D, Wang P, Shen C, Dick A, van den Hengel A. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*. 2017; 163:21–40. <https://doi.org/10.1016/j.cviu.2017.05.001>
2. Huang P, Huang J, Guo Y, Qiao M, Zhu Y. Multi-grained Attention with Object-level Grounding for Visual Question Answering. In: *Proceedings of the 57th Conference of the Association for*

- Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers; 2019. p. 3595–3600.
3. Zhao S, Sharma P, Levinboim T, Soricut R. Informative Image Captioning with External Sources of Information. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers.* 2019; p. 6485–6494.
 4. Herdade S, Kappeler A, Boakye K, Soares J. Image Captioning: Transforming Objects into Words. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada;* 2019. p. 11135–11145.
 5. Saha A, Khapra MM, Sankaranarayanan K. Towards building large scale multimodal domain-aware conversation systems. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7; 2018.* p. 696–704.
 6. Le H, Hoi S, Sahoo D, Chen N. End-to-End Multimodal Dialog Systems with Hierarchical Multimodal Attention on Video Features. In: *DSTC7 at AAAI2019 workshop;* 2019.
 7. Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JM, et al. Visual dialog. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition;* 2017. p. 326–335.
 8. Chauhan H, Firdaus M, Ekbal A, Bhattacharyya P. Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers;* 2019. p. 5437–5447.
 9. Le H, Sahoo D, Chen NF, Hoi SC. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers;* p. 5612–5623.
 10. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17. 2016;* p. 110–119.
 11. Jiang S, Ren P, Monz C, de Rijke M. Improving Neural Response Diversity with Frequency-Aware Cross-Entropy Loss. In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17. ACM;* 2019. p. 2879–2885.
 12. Agarwal S, Dšek O, Konstantas I, Rieser V. Improving Context Modelling in Multimodal Dialogue Generation. In: *Proceedings of the 11th International Conference on Natural Language Generation;* 2018. p. 129–134.
 13. Agarwal S, Dušek O, Konstantas I, Rieser V. A Knowledge-Grounded Multimodal Search-Based Conversational Agent. In: *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI;* 2018. p. 59–66.
 14. Vijayakumar AK, Cogswell M, Selvaraju RR, Sun QH, Lee S, Crandall DJ, et al. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *ArXiv.* 2016; abs/1610.02424.
 15. Ben-Younes H, Cadene R, Thome N, Cord M. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1. 2019;* p. 8102–8109.
 16. Kool W, Van Hoof H, Welling M. Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement. *Proceedings of the 36th International Conference on Machine Learning, ICML, 9-15 June 2019, Long Beach, California, USA. 2019;* 97:3499–3508.
 17. Vinyals O, Le Q. A neural conversational model. *arXiv preprint arXiv:1506.05869.* 2015.
 18. Shang L, Lu Z, Li H. 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).* vol. 1; 2015. p. 1577–1586.
 19. Sordani A, Bengio Y, Vahabi H, Lioma C, Grue Simonsen J, Nie JY. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management. ACM;* 2015. p. 553–562.
 20. Serban IV, Sordani A, Bengio Y, Courville A, Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, Phoenix, Arizona, USA; 2016.* p. 3776–3784.

21. Serban IV, Sordoni A, Lowe R, Charlin L, Pineau J, Courville AC, et al. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, February 4-9, 2017, San Francisco, California, USA; 2017. p. 3295–3301.
22. Zhang Y, Galley M, Gao J, Gan Z, Li X, Brockett C, et al. Generating informative and diverse conversational responses via adversarial information maximization. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 3-8 December 2018, Montreal, Canada; 2018. p. 1815–1825.
23. Nakamura R, Sudoh K, Yoshino K, Nakamura S. Another Diversity-Promoting Objective Function for Neural Dialogue Generation. *arXiv preprint arXiv:181108100*. 2018.
24. Xu X, Dušek O, Konstas I, Rieser V. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 31–November 4. 2018; p. 3981–3991.
25. Zhao T, Zhao R, Eskenazi M. 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, ACL 2017, Vancouver, Canada, July 30–August 4, Volume 1: Long Papers*; 2017. p. 654–664.
26. Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep reinforcement learning for dialogue generation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, Austin, Texas, USA, November 1-4; p. 1192–1202.
27. Shen X, Su H, Niu S, Demberg V. Improving variational encoder-decoders in dialogue generation. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2-7; 2018. p. 5456–5463.
28. Li J, Monroe W, Jurafsky D. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:161108562*. 2016.
29. Cao K, Clark S. Latent variable dialogue models and their diversity. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. 2017; p. 182–187.
30. Xu J, Ren X, Lin J, Sun X. DP-GAN: diversity-promoting generative adversarial network for generating informative and diversified text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 31–November 4; p. 3940–3949.
31. Gao X, Lee S, Zhang Y, Brockett C, Galley M, Gao J, et al. Jointly optimizing diversity and relevance in neural response generation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). 2019; p. 1229–1238.
32. Gao J, Bi W, Liu X, Li J, Shi S. Generating multiple diverse responses for short-text conversation. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27–February 1. vol. 33; 2019. p. 6383–6390.
33. Zou M, Li X, Liu H, Deng Z. MEMD: A Diversity-Promoting Learning Framework for Short-Text Conversation. In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, Santa Fe, New Mexico, USA, August 20-26; 2018. p. 1281–1291.
34. Chen C, Peng J, Wang F, Xu J, Wu H. Generating Multiple Diverse Responses with Multi-Mapping and Posterior Mapping Selection. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, Macao, China, August 10-16. 2019; p. 4918–4924.
35. Firdaus M, Chauhan H, Ekbal A, Bhattacharyya P. EmoSen: Generating Sentiment and Emotion Controlled Responses in a Multimodal Dialogue System. *IEEE Transactions on Affective Computing*. 2020.
36. Mostafazadeh N, Brockett C, Dolan B, Galley M, Gao J, Spithourakis GP, et al. Image-grounded conversations: Multimodal context for natural question and response generation. *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017*, Taipei, Taiwan, November 27–December 1, 2017—Volume 1: Long Papers. 2017; p. 462–472.
37. De Vries H, Strub F, Chandar S, Pietquin O, Larochelle H, Courville A. Guesswhat?! visual object discovery through multi-modal dialogue. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21-26; 2017. p. 4466–4475.
38. Gan Z, Cheng Y, Kholy AE, Li L, Liu J, Gao J. Multi-step reasoning via recurrent dual attention for visual dialog. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. 2019; p. 6463–6474.

39. Wu Q, Wang P, Shen C, Reid I, Van Den Hengel A. Are you talking to me? reasoned visual dialog generation through adversarial learning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018; 2018. p. 6106–6115.
40. Guo D, Wang H, Wang M. Dual Visual Attention Network for Visual Dialog. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019; 2019. p. 4989–4995.
41. Guo D, Wang H, Wang S, Wang M. Textual-Visual Reference-aware Attention Network for Visual Dialog. IEEE Transactions on Image Processing. 2020.
42. Alamri H, Hori C, Marks TK, Batr D, Parikh D. Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC7. In: DSTC7 at AAAI2019 Workshop. vol. 2; 2018.
43. Lin KY, Hsu CC, Chen YN, Ku LW. Entropy-Enhanced Multimodal Attention Model for Scene-Aware Dialogue Generation. In: DSTC7 at AAAI2019 workshop; 2019.
44. Liao L, Ma Y, He X, Hong R, Chua Ts. Knowledge-aware multimodal dialogue systems. In: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22–26. ACM; 2018. p. 801–809.
45. Cui C, Wang W, Song X, Huang M, Xu XS, Nie L. User Attention-guided Multimodal Dialog Systems. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25; 2019. p. 445–454.
46. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada; 2014. p. 3104–3112.
47. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. 2014; p. 1724–1734.
48. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. 2015.
49. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21. 2015; p. 1412–1421.
50. Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain; 2016. p. 289–297.
51. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural computation. 1989; 1(2):270–280. <https://doi.org/10.1162/neco.1989.1.2.270>
52. Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings. 2018.
53. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014; 15(1):1929–1958.
54. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13–15; 2010. p. 249–256.
55. Chen S, Beeferman DH, Rosenfeld R. Evaluation metrics for language models; 1998.
56. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA. Association for Computational Linguistics; 2002. p. 311–318.
57. Liu CW, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4. 2016; p. 2122–2132.
58. Niu T, Bansal M. Polite Dialogue Generation Without Parallel Data. Transactions of the Association for Computational Linguistics. 2018; 6:373–389. https://doi.org/10.1162/tacl_a_00027
59. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971; 76(5):378. <https://doi.org/10.1037/h0031619>
60. Welch BL. The generalization of student's t problem when several different population variances are involved. Biometrika. 1947; 34(1/2):28–35. <https://doi.org/10.2307/2332510>