Multimodal Dialog Systems via Capturing Context-aware Dependencies of Semantic Elements

Weidong He[1], Zhi Li[2], Dongcai Lu[3], Enhong Chen[1,2,*], Tong Xu[1,2,*],
Baoxing Huai[3], Nicholas Jing Yuan[3]
[1]Anhui Province Key Lab of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China,
[2]School of Data Science, University of Science and Technology of China, [3]HUAWEI Technologies
{hwd,zhili03}@mail.ustc.edu.cn,{ludongcai,huaibaoxing,nicholas.yuan}@huawei.com,{cheneh,tongxu}@ustc.edu.cn

ABSTRACT

Recently, multimodal dialogue systems have engaged increasing attention in several domains such as retail, travel, etc. In spite of the promising performance of pioneer works, existing studies usually focus on utterance-level semantic representations with hierarchical structures, which ignore the context-aware dependencies of multimodal semantic elements, i.e., words and images. Moreover, when integrating the visual content, they only consider images of the current turn, leaving out ones of previous turns as well as their ordinal information. To address these issues, we propose a Multimodal diAlogue systems with semanTic Elements, MATE for short. Specifically, we unfold the multimodal inputs and devise a Multimodal Element-level Encoder to obtain the semantic representation at element-level. Besides, we take into consideration all images that might be relevant to the current turn and inject the sequential characteristics of images through position encoding. Finally, we make comprehensive experiments on a public multimodal dialogue dataset in the retail domain, and improve the BLUE-4 score by 9.49, and NIST score by 1.8469 compared with state-of-the-art methods.

KEYWORDS

Multimodal Dialogue; Multimodal Semantic Elements; Multimodal Transformer

ACM Reference Format:
Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai and Nicholas Jing Yuan. 2020. Multimodal Dialog Systems via Capturing Context-aware Dependencies of Semantic Elements. In *28th ACM International Conference on Multimedia (MM'20), October 12-16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413679

1 INTRODUCTION

In recent years, we have witnessed the rise of dialogue systems and the introduction of conversational agents to the market (e.g. Apple Siri[1], Amazon Alexa[2], Microsoft Cortana[3] and Google Assistant[4]). Generally speaking, traditional dialogue systems focus on two broad categories: open-domain conversations with casual chit-chat [17, 32] and task-oriented dialogue systems that are designed to accomplish a particular task [26, 40]. However, most of the existing dialogue systems only focus on the textual or voice modality [31, 32, 39], ignoring the important visual cues. Pictures often express the intentions more vividly, and content from different modalities usually provides various complementary information, especially in the fashion domain[14, 20]. As shown in Figure 1, the user searches for a caftan by entering a query with a product picture. This facilitates the user to express demands and enables the agent to understand the products better. Therefore, there exists

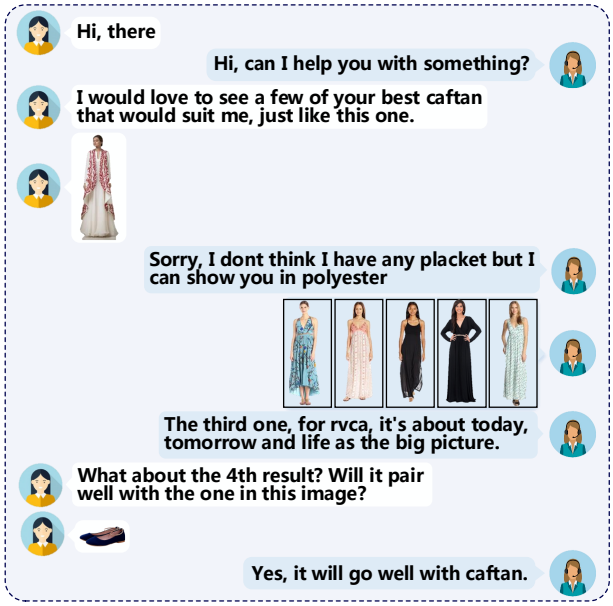

Figure 1: An example of a multimodal dialogue between a user and an agent. The user expresses her requirements and preference for products, and the agent generates the multimodal responses according to the context.

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413679

[1]https://www.apple.com/siri
[2]https://www.alexa.com
[3]https://www.microsoft.com/en-us/cortana
[4]https://assistant.google.com

an urgent requirement for conversational agents that can converse by combining different modalities.

Along this line, there are several studies integrating the visual content into the traditional textual dialogue systems, the so-called multimodal dialogue systems. As a leading study, Saha et al. [29] released a multimodal dialogue dataset (MMD) for the online retail domain. Meanwhile, they also presented a basic Multimodal HieRarchical Encoder-Decoder (MHRED) model. On the basis of MHRED model, UMD [8] devised a user attention-guided multimodal dialogue model that focuses on the user requirements in the attribute level. Nie et al. [27] employed adaptive decoders to generate general responses, knowledge-aware responses, and multimodal responses dynamically based on various user intentions.

Although existing multimodal dialogue systems have shown promising performance, they are all based on the MHRED architecture, in which the encoder compresses each utterance to a vector. For this reason, prior methods only obtain the utterance-level information, making it difficult to learn the context dependencies of multimodal semantic elements, i.e., words and images. Actually, the element-level context is essential to understand the users' intentions in the dialogue process, especially in the multimodal scenarios. For example, as illustrated in Figure 1, when the user inquires about the style-tip information, it is necessary for the agent to capture the dependencies between various multimodal semantic elements, i.e., the words "caftan", "4th" and referred images. If we simply replace "4th" by "5th", the meaning would change significantly. Merely modeling semantics at the utterance-level makes it difficult to discover these element-level semantic differences and understand the users' real intentions. Thus, in this paper, we attempt to track this problem by capturing the dependencies of multimodal semantic elements and generate context-relevance responses for dialogue systems.

However, capturing these element semantic dependencies in the multimodal dialogue system is a non-trivial task. First, exploring a unified architecture that can unfold element-level semantics is difficult. Second, when integrating the visual elements with textual context, we have to determine which images are related and take all of them from multiple conversational turns into consideration. Last, the ordinal information of images is of great importance since it is often mentioned explicitly in the dialogue, such as "4th" in the example above. How to utilize their ordinal information in dialogue context is still largely unexplored.

To address the aforementioned issues, we present a **M**ulti-modal di**A**logue system with seman**T**ic **E**lements (MATE) to deeply explore the semantic dependencies of multimodal elements in a dialogue context. To be more specific, for each turn, we first leverage a self-attention module [38] to encode textual utterances into continuous representations. Then, we use an image selector to determinate referred images from previous and current turns and extract their features using convolutional neural networks (CNN). In order to make full use of the ordinal information of these images, positional encoding is used to inject the sequential characteristics of visual features. Thereafter, an attention mechanism is used to produce joint representations of multimodal semantic elements. Finally, such representations are fed into a two-stage decoder for response generation. The first decoder concentrates on multimodal context from the encoder, while the second decoder refines the results of the first decoder by combining them with the relevant domain knowledge. This process is motivated by the human cognition process: humans usually first focus on the previous utterance, and then answer with background knowledge.

The key contributions of our work are as follows:

- We present a new perspective to address the response generation task in multimodal dialogue systems by utilizing the element-level semantics.
- We propose a novel Multi-modal diAlogue system with semanTic Elements (MATE), which is capable of capturing the dependencies of multimodal semantic elements and leveraging related images from dialogue history as well as their ordinal information to generate context-aware responses.
- We conduct extensive experiments to evaluate the proposed model and push the BLEU-4 score to 38.06 (9.49 points absolute improvement) and NIST score to 6.0604 (1.8469 points absolute improvement), compared with state-of-the-art methods. And we release our code and data[5] to facilitate the research in this field.

2 RELATED WORK

Generally, the related work of this study can be grouped into three categories: traditional dialogue systems, unimodal dialogue systems and multimodal transformer.

2.1 Unimodal Dialogue Systems

Recently, great efforts have been made to develop dialogue systems that automatically generate responses based on text or voice information. Traditional dialogue systems can be generally categorized into two groups: open-domain and task-oriented dialogue systems. The open-domain systems aim to converse with humans in diverse topics to provide reasonable responses and are usually implemented by retrieval-based or generation-based methods. Retrieval-based methods leverage the dialogue history to select proper responses from a repository, benefiting from informative and fluent responses [42, 46, 48]. By contrast, generation-based methods [17, 32] generate responses utilizing an encoder-decoder framework [36].

In contrast to former systems, task-oriented dialogue systems focus on helping users to accomplish specific tasks, such as looking for restaurants and booking movies. Traditional task-oriented dialogue systems [26, 47, 49] usually employ a typical pipeline. They first classify the users' intentions and determine users' requirements. Then, a policy network is used to decide the next action. Finally, the language generation component produces the responses through predefined templates or some generation-based models. However, such methods suffer from several serious problems [16], such as error propagation, heavy interdependence among the components and a requirement of large-scale annotated datasets.

Recently, the effectiveness of deep learning has shown remarkable improvement in dialogue systems [4, 39]. For example, Wen et al. [40] presented an end-to-end trainable dialogue system that linked input representations to slot-value pairs from a database. Serban et al. [31] extended the hierarchical encoder-decoder (HRED) neural network to generate responses. Besides, deep reinforcement learning is also used to strengthen generation-based dialogue

[5] https://github.com/githwd2016/MATE

systems [9, 18]. Although the existing systems have made much progress, they are still restricted to a single modality.

2.2 Multimodal Dialogue Systems

With the development of many industrial domains, such as travel and e-commerce retail, multimodal conversational agents are gaining importance. To this end, Saha et al. [29] constructed a Multimodal Dialogue (MMD) dataset for the fashion domain, which consists of over 150K conversation sessions and contains domain knowledge curation. Along with the dataset, they also presented a basic Multimodal HieRarchical Encoder-Decoder model (MHRED).

In order to generate reasonable responses, there are two issues that need to be considered. The first one is how to understand multimodal semantics. To this end, Liao et al. [22] extracted the visual representation using an Exclusive&Independent tree [21]. Chauhan et al. [5] proposed a novel position and attribute aware attention mechanism to learn enhanced image representation. Cui et al. [8] devised a User attention-guided Multimodal Dialogue (UMD) model that paid more attention to the user requirements explicitly in the attribute level and encoded the dialogue history dynamically based on users' attention. The second one is when and how to incorporate the domain knowledge. For this one, Liao et al. [22] stored style tips knowledge into memory networks and employed an attention mechanism to decide which knowledge entry is useful. Nie et al. [27] presented a Multimodal diAloG system with adaptIve deCoders (MAGIC), which could generate general responses, knowledge-aware responses, and multimodal responses dynamically based on various user intentions. However, all existing methods are based on MHRED, which is unable to learn the dependencies between multimodal semantic elements. Our work differs from these existing works on the MMD dataset since we propose a novel transformer-based model, which can deal effectively with the dependencies between multimodal semantic elements.

Another body of work relevant to ours is the class of Vision-to-Language problems, such as image captioning [45], visual question answering [1], video summarization [50] and cross-modal retrieval [6]. By comparison, multimodal dialogue systems focus more on multi-turn multimodal interaction between users and agents, and usually is not limited to a single image.

2.3 Multimodal Transformer

The transformer [38] was first introduced for neural machine translation (NMT) tasks and has been applied to many other tasks, such as language modeling [3] and document grounded conversations [19]. Due to the huge success of the transformer, recent works also seek to employ transformer networks for multimodal tasks, such as multimodal sentiment analysis [37] and pre-training multimodal model [24, 35]. However, how to utilize transformer networks for multimodal dialogue tasks is still unexplored. The main differences are that prior works do not maintain a multimodal context and usually use aligned data from different modalities. Nevertheless, in multimodal dialogues, we have to keep the multimodal dialogue context and select related images for each utterance. Moreover, existing works are insensitive for the order of images or only involve a single image, while in our situation an utterance might involve multiple pictures and their ordinal information is important.

3 PRELIMINARY

In this section, we first formalize our problem. Then, we introduce the Transformer Block (TB) [38], which is widely used in our model.

3.1 Problem Statement

In this paper, we focus on the task of textual response generation conditioned on multimodal conversational history as proposed in [29]. To be specific, given a multimodal dialogue history $\mathbf{H_T} = \left\{(\mathbf{U}^t, \mathbf{I}^t)\right\}_{t=1}^{T}$ and a user query $(\mathbf{U}^q, \mathbf{I}^q)$, the task is to generate the textual system response $\mathbf{U}^\mathbf{r}$. Here, $\mathbf{U}^t = \{u_i^t\}_{i=1}^{m_t}$ denotes the t-th text utterance containing m_t words, and $\mathbf{I}^t = \{i_j^t\}_{j=1}^{n_t}$ denotes the t-th image utterance containing n_t images. Note that n_t may be zero, i.e., there is no image in t-th turn. Similarly, $\mathbf{U}^q = \{u_i^q\}_{i=1}^{m_q}$ contains m_q words and $\mathbf{I}^q = \{i_j^q\}_{j=1}^{n_q}$ contains n_q images. Formally, the probability to generate the response $\mathbf{U}^\mathbf{r}$ is computed by

$$P(\mathbf{U}^\mathbf{r}|\mathbf{H}_T; \mathbf{U}^q; \mathbf{I}^q; \theta) = \prod_{i=1}^{m} P(u_i^r|\mathbf{H}_T; \mathbf{U}^q; \mathbf{I}^q; u_{<i}^r; \theta), \quad (1)$$

where $u_{<i}^r = (u_1^r, \ldots, u_{i-1}^r)$ denotes words that have been already generated and θ denotes trainable parameters.

3.2 Transformer Block

We consider two sequences $\mathbf{S}_\alpha \in \mathbb{R}^{l_\alpha \times d_\alpha}$ and $\mathbf{S}_\beta \in \mathbb{R}^{l_\beta \times d_\beta}$, where $l_{(.)}$ and $d_{(.)}$ represent sequence length and feature dimension respectively. Note that $\mathbf{S}_\alpha$ and $\mathbf{S}_\beta$ could be same, i.e., the so-called self-attention. We suppose that $\mathbf{S}_\alpha$ is the target sequence. A transformer block is composed of a stack of M identical layers. Each layer is composed of a multi-head attention sub-layer (Multihead(.)) and a position-wise fully connected feed-forward network (FFN(.)). The process is as follows:

$$\mathbf{S}_\alpha^0 = \mathbf{S}_\alpha, \quad (2)$$

$$\hat{\mathbf{S}}_\alpha^i = \text{LayerNorm}(\text{Multihead}(\mathbf{S}_\alpha^{i-1}, \mathbf{S}_\beta, \mathbf{S}_\beta) + \mathbf{S}_\alpha^{i-1}), \quad (3)$$

$$\mathbf{S}_\alpha^i = \text{LayerNorm}(\text{FFN}(\hat{\mathbf{S}}_\alpha^i) + \hat{\mathbf{S}}_\alpha^i), i = 1, \ldots, M \quad (4)$$

where LayerNorm(.) means layer normalization [2] and $\mathbf{S}_\alpha^M \in \mathbb{R}^{l_\alpha \times d_{model}}$ is the final output of the transformer block. The multi-head attention sub-layer contains h single-head attention. The input of the j-th head consists of a query matrix $\mathbf{Q}$, a key matrix $\mathbf{K}$ and a value matrix $\mathbf{V}$. The multi-head attention is as follows:

$$\mathbf{Q}_j = \mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}_j = \mathbf{K}\mathbf{W}_j^K, \mathbf{V}_j = \mathbf{V}\mathbf{W}_j^V \quad (5)$$

$$\mathbf{head}_j = \text{softmax}\left(\frac{\mathbf{Q}_j\mathbf{K}_j^\top}{\sqrt{d_k}}\right)\mathbf{V}_j, \quad (6)$$

$$\text{Multihead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{head}_1, \ldots, \mathbf{head}_h]\,\mathbf{W}^O \quad (7)$$

where $[.]$ is the concatenation operation. $\mathbf{W}_j^Q \in \mathbb{R}^{d_\alpha \times d_k}$, $\mathbf{W}_j^K \in \mathbb{R}^{d_\beta \times d_k}$, $\mathbf{W}_j^V \in \mathbb{R}^{d_\beta \times d_v}$ and $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_{model}}$ are trainable parameters. Note that in practice, we usually use $d_k = d_v = d_{model}/h$ to keep the similar computational cost as single-head attention with full dimensionality. After the multi-head attention sub-layer, a position-wise fully connected feed-forward network is injected to complete the transformer block:

$$\text{FFN}(\hat{\mathbf{S}}_\alpha^i) = \max\left(0, \hat{\mathbf{S}}_\alpha^i\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2 + \mathbf{b}_2, \quad (8)$$

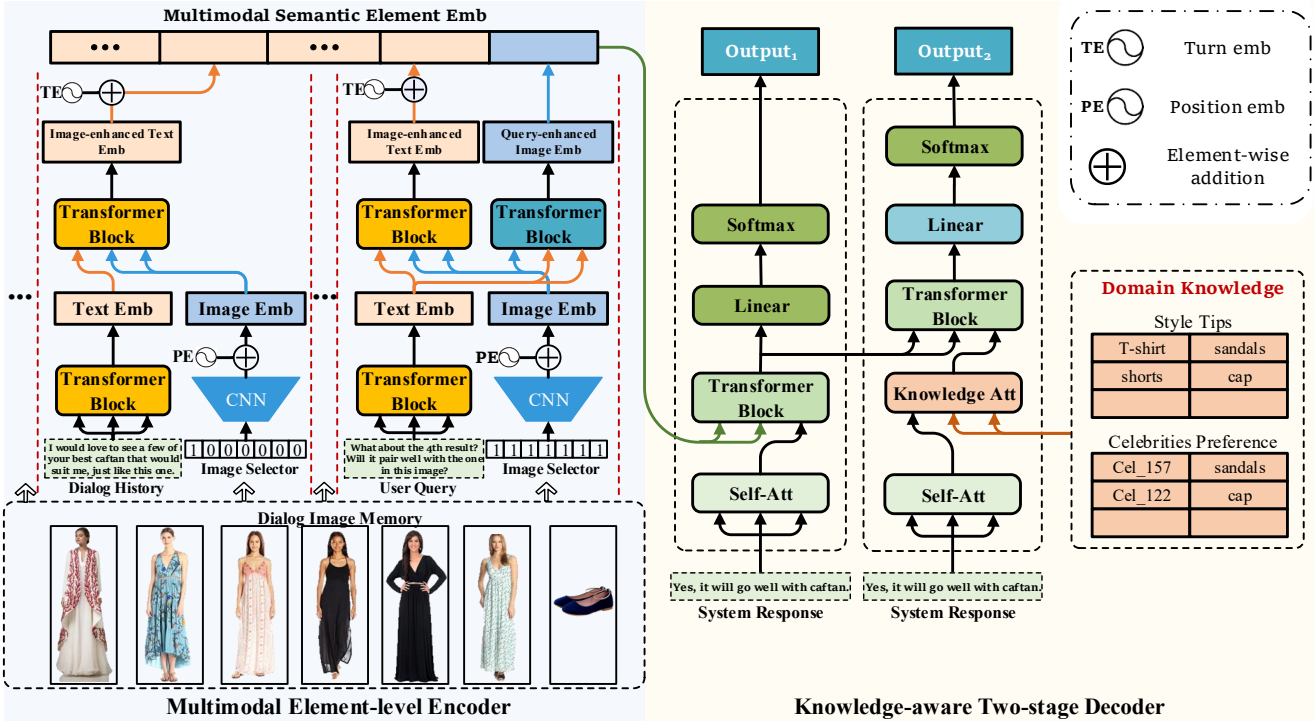

Figure 2: The framework of proposed MATE model.

4 METHODOLOGY

The overall architecture of our proposed MATE model is shown in Figure 2. It consists of two main components:

Multimodal Element-level Encoder: In this component, all images from the dialog history and the user query are organized as dialog image memory. Then, we allocate related images to each turn and obtain image-enhanced text embeddings through an attention mechanism. Meanwhile, all images are integrated with a user query to get a query-enhanced image embeddings. Finally, all embeddings are concatenated as multimodal semantic element embeddings.

Knowledge-aware Two-Stage Decoder: It is a variant of a transformer decoder for generating better responses. The first-stage decoder focuses on the multimodal conversation context from the encoder, while the second-stage decoder takes domain knowledge and results from the first decoder to further refine the responses.

4.1 Multimodal Element-level Encoder

The goal of this component is to learn joint representations of multimodal semantic elements. First, we introduce the text and image embedding. For text utterance, we utilize the mentioned transformer block (TB) to obtain its representation. Formally, the text embedding of t-th turn $\mathbf{E}^t \in \mathbb{R}^{m_t \times d_{model}}$ is calculated by

$$\mathbf{e}_i^t = u_i^t \mathbf{W}_{emb} + \text{PE}(i), \tag{9}$$

$$\mathbf{In}^t = \left[\mathbf{e}_1^t, \ldots, \mathbf{e}_{m_t}^t\right], \tag{10}$$

$$\mathbf{E}^t = \text{TB}(\mathbf{In}^t, \mathbf{In}^t, \mathbf{In}^t)), \tag{11}$$

where $\mathbf{W}_{emb} \in \mathbb{R}^{d_{vocab} \times d_{model}}$ is the word embedding matrix, and d_{vocab} is the size of vocabulary. PE(.) is the positional embedding [38] to make use of the order of the sentence.

As for the visual modality, we utilize convolutional neural networks (CNN), such as VGGNet-16 [33] or ResNet [13], to obtain initial image embedding. Compared with the previous works, our method has two main differences. The first one is that we not only employ images in the current turn, but also previous images. In our opinion, it's important to take previous images into consideration, especially for the turn without any pictures. To be specific, for each turn, we utilize an Image Selector to mask out images after the current turn. The second one is that we also construct position encoding for images since the ordinal information of the images is essential for generating the correct system responses. Formally, we denote image memory for turn t as $\mathbf{Img} = (i_1^1, \ldots, i_{n_1}^1, \ldots, i_1^t, \ldots, i_{n_t}^t)$, and corresponding image representation as $\mathbf{V}^t \in \mathbb{R}^{N_t \times d_{image}}$, where $N_t = \sum_{i=1}^{t} n_i$ and d_{image} is the embedding size of images. The image embedding of t-th turn $\mathbf{V}^t$ is calculated by

$$\mathbf{v}_i^t = \text{CNN}(\mathbf{Img}_i) + \text{PE}(i), \tag{12}$$

$$\mathbf{V}^t = \left[\mathbf{v}_1^t, \ldots, \mathbf{v}_{N_t}^t\right] \tag{13}$$

Afterwards, we introduce the information fusion of two modalities (on the left-top of Figure 2). For text utterance in turn t, we utilize the corresponding image embeddings to enhance the text representations. Besides, in order to indicate the dialogue turn explicitly, we devise to add turn embeddings (TE(.)) to text embeddings of each turn, which are similar to positional embeddings. The t-th image-enhanced text embeddings $\tilde{\mathbf{E}}^t$ is calculated by

$$\tilde{\mathbf{E}}^t = \text{TB}(\mathbf{E}^t, \mathbf{V}^t, \mathbf{V}^t) + \text{TE}(t), \tag{14}$$

The situation is similar for the user query. The main difference is that we also utilize the user query to enhance dialog images.

The motivation is that user queries often provide clues about the image importance of responses. Finally, we concatenate the image-enhanced text embeddings and query-enhanced image embeddings to obtain the context embeddings, which we call multimodal semantic element embeddings. The process is formulated as follows:

$$\tilde{\mathbf{E}}^q = \text{TB}(\mathbf{E}^q, \mathbf{V}, \mathbf{V}) + \text{TE}(T+1), \tag{15}$$

$$\tilde{\mathbf{V}}^q = \text{TB}(\mathbf{V}, \mathbf{E}^q, \mathbf{E}^q), \tag{16}$$

$$\mathbf{C} = \left[\tilde{\mathbf{E}}^1, \ldots, \tilde{\mathbf{E}}^T, \tilde{\mathbf{E}}^q, \tilde{\mathbf{V}}^q\right], \tag{17}$$

where $\mathbf{E}^q$ is text embeddings of user query, which are obtained like $\mathbf{E}^t$. $\mathbf{V}$ is the representations of all images in memory and $\mathbf{C} \in \mathbb{R}^{N_c \times d_{model}}$ is the final output of multimodal encoder, where $N_c = \sum_{i=1}^{T} m_i + m_q + \sum_{i=1}^{T} n_i + n_q$. It is worth noticing that in practice, we do not use the complete conversation history because of storage and computation power limitations.

4.2 Knowledge-aware Two-Stage Decoder

This component utilizes multimodal context information from the encoder and structured domain knowledge to generate context- and knowledge-aware system responses. Inspired by [19], it consists of a first-stage decoder and a second-stage decoder. The first-stage decoder takes the multimodal semantic elements as input and focuses on how to generate responses contextual coherently. The second-stage decoder takes the first-stage decoding results and domain knowledge as input and focuses on increasing knowledge usage and guiding the following conversations. When generating the i-th word $\hat{u}_i^r$, we have already generated the previous words $\hat{u}_{<i}^r$. We use $\mathbf{In}^r$ to denote the embeddings of those earlier generated words as following:

$$\mathbf{e}_i^r = \hat{u}_i^r \mathbf{W}_{emb} + \text{PE}(i), \tag{18}$$

$$\mathbf{In}^r = \left[\mathbf{e}_{SOS}^r, \mathbf{e}_1^r, \ldots, \mathbf{e}_{i-1}^r\right], \tag{19}$$

where $\mathbf{e}_{SOS}^r$ is the vector representation of the sentence-start token.

As shown in Figure 2 (right), two decoders have a similar architecture but different input for sub-layers. The first-stage decoder is identical to the original transformer decoder:

$$\mathbf{E}_1^r = \text{MultiHead}(\mathbf{In}_1^r, \mathbf{In}_1^r, \mathbf{In}_1^r), \tag{20}$$

$$\mathbf{R}_1 = \text{TB}(\mathbf{E}_1^r, \mathbf{C}, \mathbf{C}), \tag{21}$$

$$P(\hat{u}_{1,<(i+1)}^r) = \text{softmax}(\mathbf{R}_1 \mathbf{W}_1^{proj}), \tag{22}$$

where $\mathbf{In}_1^r$ is the input for first-stage decoder calculated by Equation 19, and $\mathbf{E}_1^r$ is the embeddings of generated words after the self-attention layer. In addition, $\hat{u}_{1,<(i+1)}^r = (\hat{u}_{1,1}^r, \ldots, \hat{u}_{1,i}^r)$ are the words decoded by the first-stage decoder, and $\mathbf{W}_1^{proj} \in \mathbb{R}^{d_{model} \times d_{vocab}}$ is the linear projection matrix.

For the second-stage decoder, we first introduce the related domain knowledge. Inspired by Nie et al. [27], in the MMD dataset, we introduce two kinds of domain knowledge, namely style tips and celebrities preference. Specifically, style tips describe the match between different products, such as sandals going well with T-shirts, while celebrities preference presents the preference distribution of celebrities over products. For example, some celebrities prefer black T-shirts over blue ones.

To incorporate the style tips knowledge, take *(T-shirts, sandals)* as an example, first we embed *T-shirts* and *sandals* into vectors respectively and then concatenate them to obtain the knowledge entry. Thus, we obtain the style tips matrix $\mathbf{ST} \in \mathbb{R}^{N_s \times d_{kng}}$, where N_s is the number of style tips and d_{kng} is the embedding size of knowledge. Similarly, for celebrities preference, we embed celebrities and products separately and then concatenate them. We use $\mathbf{CP} \in \mathbb{R}^{N_c \times d_{kng}}$ to denote knowledge entries of celebrities preference, where N_c is the number of celebrities. Finally, the representation of domain knowledge $\mathbf{DK} \in \mathbb{R}^{(N_s+N_c) \times d_{kng}}$ is obtained by concatenating $\mathbf{ST}$ and $\mathbf{CP}$. The second-stage decoder generates responses as follows:

$$\mathbf{E}_2^r = \text{MultiHead}(\mathbf{In}_2^r, \mathbf{In}_2^r, \mathbf{In}_2^r), \tag{23}$$

$$\mathbf{H} = \text{MultiHead}(\mathbf{E}_2^r, \mathbf{DK}, \mathbf{DK}), \tag{24}$$

$$\mathbf{R}_2 = \text{TB}(\mathbf{H}, \mathbf{R}_1, \mathbf{R}_1), \tag{25}$$

$$P(\hat{u}_{2,<(i+1)}^r) = \text{softmax}(\mathbf{R}_2 \mathbf{W}_2^{proj}), \tag{26}$$

where $\mathbf{H}$ is the hidden state from the knowledge attention layer, and $\hat{u}_{2,<(i+1)}^r$ are the words produced by the second-stage decoder.

4.3 Model Training

For training, we employ the commonly used teacher forcing [41] algorithm at every decoding step. Our two-stage decoder is inspired by Deliberation Network [43]. In the original paper, they proposed a complex joint learning framework to train the model. In contrast to them, we minimize the negative log-likelihood loss from two decoders, following Xiong et al. [44]:

$$L_{mle} = L_{mle1} + L_{mle2}, \tag{27}$$

$$L_{mle1} = -\sum_{k=1}^{K} \sum_{i=1}^{m_r^k} \log P(u_{1,i}^r), \tag{28}$$

$$L_{mle2} = -\sum_{k=1}^{K} \sum_{i=1}^{m_r^k} \log P(u_{2,i}^r), \tag{29}$$

where L_{mle1} and L_{mle2} are the loss from the first and second stage decoder, respectively. K is the number of responses in the dataset and m_r^k is the number of words in k-th responses.

5 EXPERIMENTS

In this section, we conduct external experiments to evaluate our proposed method on a real-world dataset. We first introduce the experimental dataset and settings, including hyper parameters, evaluation metrics and compared methods. Then, we present the experimental results and analyses from multiple perspectives to answer the following research questions:

(1) Can our model generate better responses compared with state-of-the-art methods?
(2) What are the effects of employing dialog image memory and image position embedding in our model?
(3) Will element-level semantic embeddings help to improve the performance of response generation in dialogue systems?

5.1 Dataset

We utilize the Multimodal Dialogue (MMD) dataset from [29] in the retail domain. During the conversations, customers mention their requirements and the agent introduces different products step by step until they make a deal. The dialogues seamlessly incorporate multimodal data in utterances and also demonstrate domain-specific knowledge during the conversation. Over 1 million fashion products with their available semi/unstructured information are collected from several well-known online retailing websites, such as Amazon[6], Jabong[7], and Abof[8].

Based on the MMD benchmark dataset, Saha et al. [29] proposed two major research tasks, including the textual response generation and the image response selection task. The former task is to generate the next text response when given a context of k turns. The latter task is to retrieve and rank m images from a database based on their relevance to the given context. Our work focuses on the textual response generation task.

5.2 Experiment Setup

We implemented our model using the deep learning framework PyTorch[9]. Following former studies [8, 27, 29], we use two-turn utterances before the responses as the context in the training period. The vocabulary size is 26,422 and the low-frequency words in the vocabulary are mapped to the special token "UNK". The dimension of word embeddings is 512, which was determined empirically. This dimension is shared by utterances and generated responses. The number of layers of both encoder and decoder are set to 3. The number of attention heads in the multi-head attention is 8 and the inner-layer size is 2048, as described in [38]. We use a dropout rate of 0.1 [34]. The model parameters are randomly initialized using a Gaussian distribution with Xavier scheme [12]. We use Adam [15] for optimization and the initial learning rate is 1e-5.

5.3 Compared Methods

To demonstrate the effectiveness of our proposed model, we compare it with the following representative methods:

Seq2seq: It is a classic encoder-decoder framework [36] with global attention [25] that has demonstrated its effectiveness in many natural language processing tasks.

HRED: HRED is the most representative method [30] in textual multi-turn dialogue systems. It is composed of a word-level LSTM for each sentence and a sentence-level LSTM connecting utterances.

MHRED: The multimodal hierarchical encoder-decoder from Saha et al. [29] incorporates the visual features into the basic HRED model and achieves a promising performance. This is the first work on multimodal task-oriented dialogue systems in the retail domain.

UMD: Based on MHRED, the user attention-guided multimodal dialogue system by Cui et al. [8] considers the hierarchical product taxonomy and the users' attention to products.

OAM: Chauhan et al. [5] proposes a novel ordinal and attribute aware attention mechanism for natural language generation exploiting images and texts. We refer to this model as OAM.

MAGIC: Multimodal dialog system with adaptive decoders [27] leverages user intentions explicitly to generate general responses, knowledge-aware responses, and multimodal responses dynamically. It's the strongest baseline that achieves the best performance on the MMD dataset.

5.4 Evaluation Metrics

To understand the quality of responses, we adopt both automatic and human evaluation methods to compare the performance of different models.

5.4.1 ***Automatic Evaluation.*** We use the BLEU-N [28], and NIST [10] as automatic evaluation metrics, following recent studies in this field [27, 29]. Since the length of about 20% target responses in the MMD dataset is less than 4, we calculate BLEU-N by varying N from 1 to 4. Higher Bleu scores mean that more n-gram overlaps between the predicted and target responses. Based on BLEU, NIST considers the weights of n-grams dynamically, i.e., the weight of an n-gram is proportional to its rareness. We use the same evaluation scripts[10] as MAGIC [27].

5.4.2 ***Human Evaluation.*** Considering that the automatic metrics are not always completely accurate to evaluate the responses [23], we also evaluate the dialogue generation based on the opinion of humans. We randomly sample 15 conversations containing 300 multimodal contexts from the testing data and then feed these contexts into MATE and two state-of-the-art models (OAM and MAGIC) to generate the textual responses. Thereafter, the 300 responses of MATE are compared with the corresponding responses generated by the two baselines. In this way, we obtain 600 pair-wise responses. After that, three experts are asked to compare the pair-wise responses from three perspectives independently:

- Fluency: Whether the generated response is grammatically correct, natural, and fluent.
- Context Coherence: Whether the response is in accordance with the aspect being discussed (style, colour, etc.) and guides the following dialogue utterances.
- Knowledge Relevance: Whether the response uses relevant and correct domain knowledge, such as celebrities preference and style tips.

The annotators are invited to judge which response is better in the context. If two responses are both meaningful or inappropriate, the comparison of this pair is treated as "draw". Ultimately, we average the results of three experts and calculate their Fleiss' kappa scores [11].

5.5 Experimental Results

In this section, we present the detailed experimental results using automatic and human evaluation metrics simultaneously.

5.5.1 ***Automatic Evaluation.*** Table 1 shows the results of automatic evaluation between baselines and MATE. From that, we have the following observations. First, MATE significantly outperforms the baselines on both BLEU and NIST scores. For example, in terms of BLEU-4 score, it improves the performance by 10.64 and 9.49 points as compared to the state-of-the-art model OAM and MAGIC,

[6]https://www.amazon.com/
[7]https://www.jabong.com
[8]https://www.abof.com
[9]https://pytorch.org/
[10]https://www.nist.gov/itl/iad/mig/tools

Table 1: Performance comparison between the different models on textual response generation. Results with † are reported by Nie et al. [27].

Methods		BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST
Text-only	**Seq2seq**† (Sutskever et al. [36])	35.39	28.15	23.81	20.65	3.3261
	HRED† (Serban et al. [30])	35.44	26.09	20.81	17.27	3.1007
Multimodal	**MHRED**† (Saha et al. [29])	32.60	25.14	23.21	20.52	3.0901
	UMD (Cui et al. [8])	44.97	35.06	29.22	25.03	3.9831
	OAM (Chauhan et al. [5])	48.30	38.24	32.03	27.42	4.3236
	MAGIC† (Nie et al. [27])	50.71	39.57	33.15	28.57	4.2135
	MATE (first-stage)	**56.08**	**47.47**	**41.86**	**37.65**	**6.0037**
	MATE (second-stage)	**56.55**	**47.89**	**42.48**	**38.06**	**6.0604**

Table 2: Human evaluation results between our model and other baselines regarding three evaluation factors.

	Fluency				Context Coherence				Knowledge Relevance			
Opponent	Win	Loss	Draw	Kappa	Win	Loss	Draw	Kappa	Win	Loss	Draw	Kappa
vs. OAM	**33.6%**	13.8%	52.6%	0.70	**59.9%**	16.4%	23.7%	0.40	**66.4%**	11.8%	21.7%	0.51
vs. MAGIC	**25.8%**	15.2%	59.0%	0.65	**49.6%**	15.6%	34.8%	0.51	**48.8%**	21.3%	29.9%	0.53

respectively. The superior performance of the proposed method demonstrates the usefulness of the novel architecture. Secondly, the BLEU-1 score of our model is relatively high. When analyzing the MMD dataset, we find that there are a lot of *"Yes or No"* responses (around 12% of total responses) about knowledge-aware questions, e.g., *"Does T-shirts go well with sandals?"*. We calculate the accuracy of these responses and our model achieves superior accuracy up to 90.14%. Thirdly, comparing results from the first and second stage decoder, we find that MATE benefits from the two-stage setting. A more detailed analysis of individual model components is provided in an ablation study later.

5.5.2 ***Human Evaluation.*** Table 2 illustrates the human evaluation results. Firstly, the kappa scores indicate a substantial agreement on fluency, and a moderate agreement on context coherence and knowledge relevance among the annotators. Secondly, MATE surpasses the baselines in all comparisons, especially in the context coherence and knowledge relevance. It demonstrates that MATE is capable of utilizing element-level context information and domain knowledge. Thirdly, all three models perform relatively well on fluency so that the annotators often assign equal ratings.

5.6 Model Ablation

Although our model shows good performance, the contributions of each model components are unclear. Hence, we conduct ablation study and Table 3 lists the results. First, we evaluate the effectiveness of introducing dialog image memory. Compared with model 1, model 2 and 3 remove the positional embedding and previous images, respectively. We can see that model 1 outperforms the two variants, which demonstrates the necessity of the dialog image memory. Then, we test the influence of the multimodal semantic elements. To be specific, for model 4, we utilize a bidirectional Gate Recurrent Units (GRU) [7] to encode each image-enhanced text from our multimodal encoder and discard the query-enhanced image embeddings. The final hidden states of each GRU are concatenated and sent to the decoder. Through this way, the decoder only receives the utterance-level information, just like MHRED. We can observe that the performance drops significantly, demonstrating the importance of introducing multimodal semantic elements.

Table 3: Ablation study of the proposed model.

Methods	No.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST
MATE	1	**56.55**	**47.89**	**42.48**	**38.06**	**6.0604**
- image position	2	55.50	47.27	41.79	37.65	5.9744
- previous images	3	55.01	46.69	41.21	37.10	5.8711
- element-level	4	54.04	45.69	40.23	36.10	5.7407
- turn embedding	5	55.03	46.72	41.21	37.03	5.8765
- knowledge	6	55.33	46.96	41.47	37.30	5.9453
ME + HRED	7	52.51	43.24	37.38	32.96	5.1820
MHRED	8	32.60	25.14	23.21	20.52	3.0901

Moreover, we also eliminate turn embedding (model 5). The result shows that the impact of turn embedding on the final performance is almost the same as the dialog image memory. Thereafter, we remove the second-stage decoder to test the effect of knowledge (model 6). However, the effect of domain knowledge is not as obvious as in previous works [27]. The possible reason is that we did not assign a suitable type of domain knowledge based on users' intentions classification. For example, introducing knowledge into general responses like "*Hi, can I help you?*" might have a negative impact. We plan to investigate the effective use of domain knowledge in our future work.

Last but not least, we also test the entire multimodal encoder. In particular, we replace the low-level encoder in MHRED by our multimodal encoder (model 7). We surprisingly discover that it improves the performance of HRED a lot, it even surpasses MAGIC. This result shows the effectiveness of our encoder.

5.7 Case Study

Figure 3 lists four typical responses sampled from the testing data. Due to the limited space, we only show the responses generated by OAM, MAGIC and MATE, and omit the part context with an ellipsis. From these cases, we have several observations. First, for the general responses, we find that all three models generate appropriate

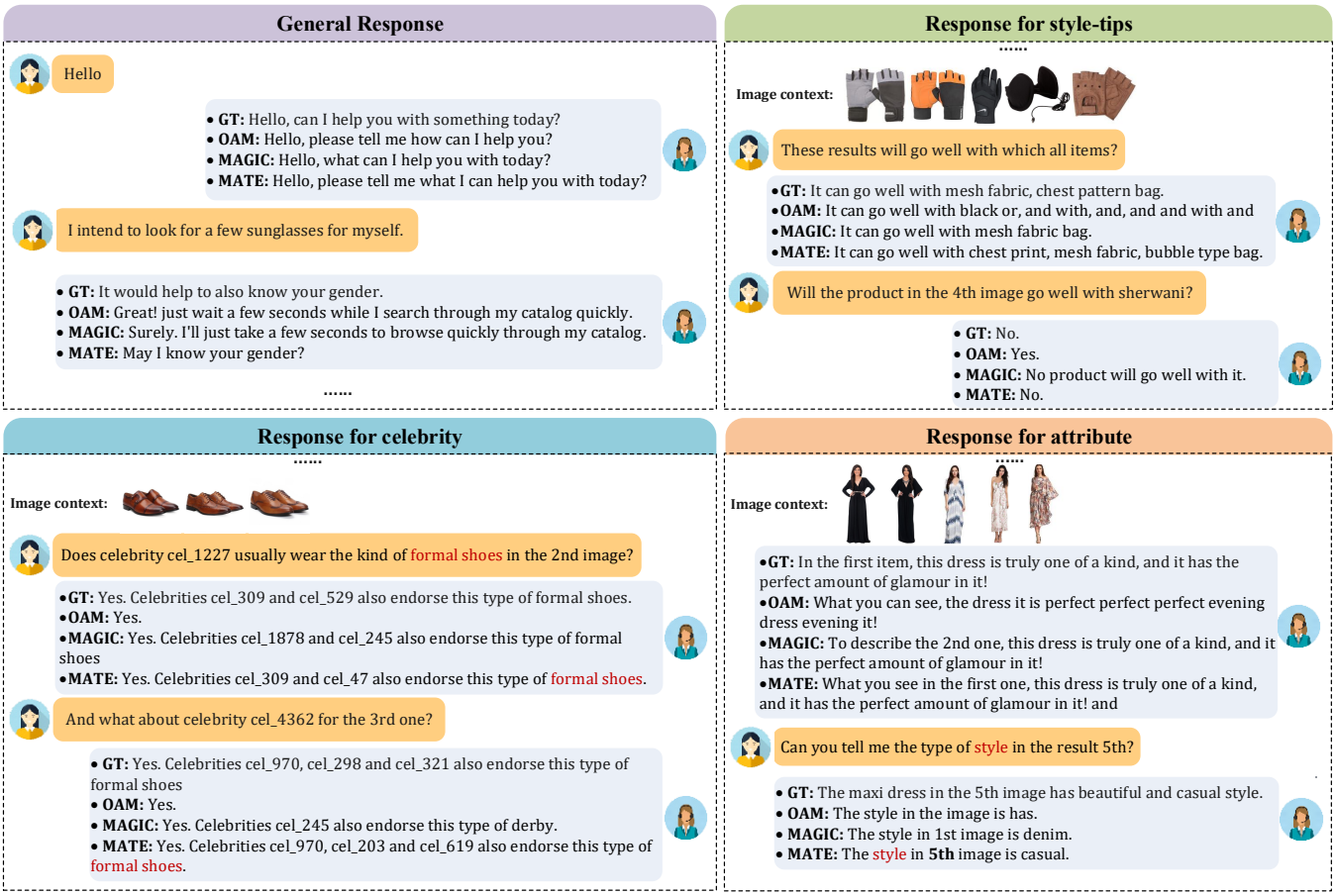

Figure 3: Four typical responses generated by the different models, including the ground truth responses (GT), and the responses generated by OAM, MAGIC and MATE.

responses. The only thing worth noting is that MATE performs better when requiring users to provide personal information (see second turn response in the general response case). Then, when answering questions about style-tips, responses generated by our model are more accurate. In the first turn, OAM fails to generate meaningful responses, while the other two succeed. Moreover, our model produces more informative sentences. As for the celebrity, we discover that it is a relatively difficult task. In the case, we could find that all three models fail to predict all celebrities correctly. In most cases, only partial celebrities are predicted correctly, or only "*Yes / No*" are answered. Finally, we discuss the responses about attributes. In the first turn, both MAGIC and MATE give perfect descriptions, while the former chooses the wrong image. Thanks to the incorporation of image position embedding, our model picks the right image. Meanwhile, our model also captures important semantic elements in the context, such as "formal shoes" in the celebrity case and "style" and "5th" in the attribute case. This ability is crucial for the agent to produce context-aware answers.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel multimodal dialogue system with semantic elements for natural language generation in multimodal dialogue systems. In particular, we first unfolded the multimodal dialogue context and utilized a multimodal element-level encoder for effective integration. Thereafter, we leveraged a knowledge-aware two-stage decoder for response generation, consisting of two decoders that could deal with context information and domain knowledge respectively. Extensive experiments exhibited the superiority of our proposed model on response generation, demonstrating the effectiveness of the architecture.

Our study may bring some new insights for unfolding the element-level semantics to multimodal dialogue systems. In the future, we will further explore the use of domain knowledge, especially the cross-modal knowledge, such as the visual patterns of style. Moreover, we will extend and apply our model in some other tasks, such as the image response selections.

7 ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (No.61727809, U1605251, 61703386), and the grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600). This work was also partially supported by the HUAWEI-USTC Joint Innovative Project and Language and Speech Innovation Lab of HUAWEI Cloud.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Alexei Baevski and Michael Auli. 2019. Adaptive Input Representations for Neural Language Modeling. In *International Conference on Learning Representations*.

[4] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning End-to-End Goal-Oriented Dialog. In *International Conference on Learning Representations*.

[5] Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5437–5447.

[6] Joya Chen, Hao Du, Yufei Wu, Tong Xu, and Enhong Chen. 2020. Cross-modal video moment retrieval based on visual-textual relationship alignment. *SCIENTIA SINICA Informationis* 50, 6 (2020), 862–876.

[7] Kyunghyun Cho, Bart van Merriënboer, 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)*. 1724–1734.

[8] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User Attention-guided Multimodal Dialog Systems. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 445–454.

[9] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmad, and Li Deng. 2017. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 484–495.

[10] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*. 138–145.

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

[12] 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*. 249–256.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. 2019. Explainable fashion recommendation: a semantic attribute region guided approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 4681–4688.

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[16] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1437–1447.

[17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 110–119.

[18] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1192–1202.

[19] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 12–21.

[20] Zhi Li, Bo Wu, Qi Liu, Likang Wu, Hongke Zhao, and Tao Mei. 2020. Learning the Compositional Visual Coherence for Complementary Recommendations. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. AAAI Press, 3536–3543.

[21] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia*. 1571–1579.

[22] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*. 801–809.

[23] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2122–2132.

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[25] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1412–1421.

[26] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1777–1788.

[27] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal Dialog System: Generating Responses via Adaptive Decoders. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1098–1106.

[28] 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*. Association for Computational Linguistics, 311–318.

[29] 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*.

[30] 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 (2015).

[31] 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 *Thirtieth AAAI Conference on Artificial Intelligence*.

[32] 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)*. Association for Computational Linguistics, 1577–1586.

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

[34] 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 (2014), 1929–1958.

[35] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

[36] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[37] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6558–6569.

[38] 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*. 5998–6008.

[39] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[40] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 438–449.

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

[42] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 496–505.

[43] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*. 1784–1794.

[44] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7338–7345.

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on*

machine learning. 2048–2057.

[46] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 55–64.

[47] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[48] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 245–254.

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

[50] Peilun Zhou, Tong Xu, Zhizhuo Yin, Dong Liu, Enhong Chen, Guangyi Lv, and Changliang Li. 2019. Character-oriented Video Summarization with Visual and Textual Cues. *IEEE Transactions on Multimedia* (2019).