Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems

Zheng Zhang[1,+], Lizi Liao[2,+], Minlie Huang[1,*], Xiaoyan Zhu[1], Tat-Seng Chua[2]
[1]Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems
[1]Beijing National Research Center for Information Science and Technology
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]National University of Singapore
{zhangz.goal,liaolizi.llz}@gmail.com, {aihuang,zxy-dcs}@tsinghua.edu.cn, chuats@comp.nus.edu.sg

ABSTRACT

Multimodal dialogue systems are attracting increasing attention with a more natural and informative way for human-computer interaction. As one of its core components, the belief tracker estimates the user's goal at each step of the dialogue and provides a direct way to validate the ability of dialogue understanding. However, existing studies on belief trackers are largely limited to textual modality, which cannot be easily extended to capture the rich semantics in multimodal systems such as those with product images. For example, in fashion domain, the visual appearance of clothes play a crucial role in understanding the user's intention. In this case, the existing belief trackers may fail to generate accurate belief states for a multimodal dialogue system.

In this paper, we present the first neural multimodal belief tracker (**NMBT**) to demonstrate how multimodal evidence can facilitate semantic understanding and dialogue state tracking. Given the multimodal inputs, while applying a textual encoder to represent textual utterances, the model gives special consideration to the semantics revealed in visual modality. It learns concept level fashion semantics by delving deep into image sub-regions and integrating concept probabilities via multiple instance learning. Then in each turn, an adaptive attention mechanism learns to automatically emphasize on different evidence sources of both visual and textual modalities for more accurate dialogue state prediction. We perform extensive evaluation on a multi-turn task-oriented dialogue dataset in fashion domain and the results show that our method achieves superior performance as compared to a wide range of baselines.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Multimedia and multimodal retrieval**; *Users and interactive retrieval*;

ACM Reference Format:
Zheng Zhang[1,+], Lizi Liao[2,+], Minlie Huang[1,*], Xiaoyan Zhu[1], Tat-Seng Chua[2]. 2019. Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems. In *Proceedings of The Web Conference 2019 (TheWebConf 2019).* ACM, New York, NY, USA, 11 pages.

TheWebConf 2019, May 13–17, 2019, San Francisco, USA

1 INTRODUCTION

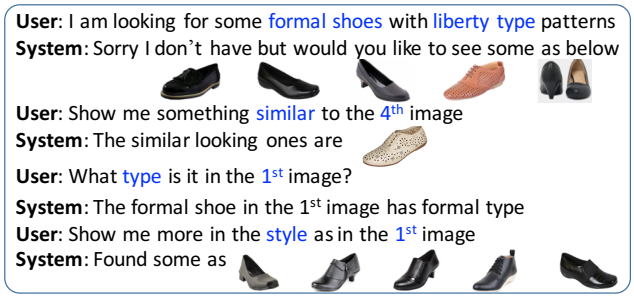

Figure 1: An illustrative example of multimodal dialogue for fashion retail, which demonstrates the importance of visual modality in understanding the inherent semantics. Keywords in blue correspond to images, where both modalities are crucial for dialogue state tracking.

By offering a natural and interactive way to satisfy user's information need, multimodal dialogue systems [34] have attracted more and more attention recently. Compared to traditional text-based systems, multimodal dialogue systems enable users to easily provide an image sample instead of racking their minds for an appropriate text description, such as in search of fashion products. At the same time, it is more straight-forward for users to perceive information from system provided images rather than text based on supposition. In task oriented scenarios, multimodal dialogue systems can better help users achieve their goals such as finding specific fashion products or travel sights under the help of visual modality and has achieved superior performance [26].

Efficient operation of such dialogue systems requires a core component — belief tracker (or known as dialogue state tracking) that can track what has happened by modeling system outputs, user utterances, and context from previous turns *etc.* A belief tracker provides a direct way to validate the systems' understanding of user's goal at each step of the dialogue. At the same time, the output of the belief tracker also supports the downstream dialogue policy component to decide what action the system should take next. Such output offers an explicit way to evaluate whether the learned policy is reasonable. Therefore, the belief tracker component is essential for the final performance of a complete dialogue system.

* Corresponding author.
+ These two authors contributed equally.

However, since current dialogue systems are largely confined to textual modality, existing efforts on belief tracker are also limited primarily to text-based methods. Thus, it may fail to generate accurate belief states for multimodal systems, as the rich semantics inherent in visual modality are ignored. For example, in order to generate correct belief states for each turn of the dialogue as shown in Figure 1, the system needs to understand the visual features of the fourth image, the type and style of the first image, *etc.* Purely relying on text, the system would fail to obtain these useful semantics, and thus result in inaccurate belief states. Moreover, it may further affect the performance of other downstream dialogue system components. Therefore, the primary goal of our work is to build a multimodal belief tracker which is able to accurately understand multimodal evidence and adaptively integrate these modalities for dialogue state tracking.

Although we have shown that multimodal evidence is crucial for dialogue state tracking, it is non-trivial to correctly extract useful semantics and leverage them appropriately in dialogue state tracking. **First**, there exists the well-known semantic gap between the low-level visual cues and the high-level semantics (*e.g., style, fit*). Indeed, there are many studies focusing on the Vision-to-Language problems such as image captioning [27], visual storytelling [18] and visual question answering (VQA) [2], which have achieved good performance. Still, as pointed out in [10, 45], there exist strong language priors which lead to good superficial performance of these models without truly understanding the visual content. For example, for questions starting with "Do you see a ..." in VQA dataset, blindly answering "yes" can achieve 87% accuracy. **Second**, the heterogeneity between the visual and textual modalities makes it hard to incorporate them well to generate concrete belief states. Due to the difference in expressiveness, users tend to prefer image for illustrating certain concepts (*such as style, pattern*) while use text for describing others (*such as category, material*). Finding out these patterns of user behavior is essential for improving dialogue state tracking.

In this paper, we propose a neural multimodal belief tracker (**NMBT**) (see Figure 3) and apply it for semantic understanding across modalities and user intent tracking. First, to fill the semantic gap and avoid being misled by language priors, we develop a visual concept learner in a weakly-supervised manner by leveraging multiple instance learning [5] which reasons with sub-regions rather than the full image. The visual concept learner enables the tracker to understand fine-grained semantics of product images at the concept level. Second, to model user's behavior patterns regarding different modalities, the tracker employs an attention mechanism to adaptively attend to images and texts during dialogue state tracking at each turn. Conditioned on both textual contexts and visual contents observed so far, the model automatically learns to emphasize the evidence coming from user provided image, system provided image, or the text utterances for more accurate dialogue state tracking.

To the best of our knowledge, this is the first work for dialogue state tracking or belief tracking in multimodal dialogue systems. The main contributions for this work are as follows:

- We validate the importance of integrating multimodal evidence in dialogue state tracking and identify the critical challenges in understanding as well as leveraging such evidence.
- We delve into image sub-regions to learn concept level visual semantics and propose an adaptive attention mechanism for automatically deciding the evidence source for dialogue state tracking based on multimodal dialogue context.
- We conduct extensive experiments to evaluate the proposed method in various evaluation metrics and show superior performance over state-of-the-art methods.

2 RELATED WORK

2.1 Text-based Dialogue State Tracking

Since spoken interaction promises a natural, effective, and hands- and eyes-free method for human-computer interaction, together with the progress in natural language processing, dialogue systems have been mainly developed within textual modality. In this paper we focus on the dialogue state tracking (DST) task in task-oriented dialogue systems. Here, we summarize recent work on DST and discuss the major difference of our work.

Early dialogue systems used hand-crafted rules for DST, keeping track of a single top hypothesis for each slot of the belief state [23, 46]. Such systems require no training data and allow developers to incorporate domain knowledge to boost performance. However, such methods fail to make use of the entire N-best hypothesis list, thus do not account for uncertainty in a principled way. In addition, uncertainty also arises from the inherent ambiguity of natural language.

Therefore, statistical DST methods are introduced to better solve the uncertainty problem. Generally speaking, the statistical methods can be categorized into two types, namely generative and discriminative approaches [12]. Generative approaches maintain a distribution over dialogue states, which is calculated by Bayes rules based on the history of observations (NLU results and previous system action) in each turn. Early generative approaches attempt to formalize the dialogue system as a Markov decision process (MDP) [24, 25, 44].However, MDP models have an assumption that the state is observable, which cannot account for the uncertainty in either dialogue state or user act parsed by NLU. Therefore, [43] further proposed POMDP-based dialogue model, treating the dialogue state as hidden variable which can be inferred from system observations, and achieved better performance. However, since all possible states are enumerated in the above methods, it can be intractable when the state space is very large [12]. Also, as pointed out by [40], generative approaches must model all the correlations in the input features, so they cannot easily exploit arbitrary but potentially useful features.

Discriminative approches have the key benefits that they can incorporate a large number of features, and can be optimized directly for prediction accuracy. [3] proposed the first discriminative state tracking trained from data where features were taken from spoken language understanding (SLU) output and dialogue history. Subsequent work has explored numerous variations of this approach. For instance, [41] applied a ranking algorithm which has the ability to construct conjunctions of features. [15] applied a deep neural network as a classifier. More recently, [16] proposed an RNN-based DST to directly take automatic speech recognition results as input without extra semantic understanding and obtains better performance than those which only utilizes semantic features produced

by external NLU. [30] proposed a CNN-based method to extract semantic features from raw ASR results and passed it to DST.

As observed in the dialogue state tracking challenges (DSTC), discriminative methods tend to dominate all other approaches [39]. Our work is also based on the discriminative statistical DST framework. However, existing approaches are constrained within textual modality while ignores the rich semantics inherent in visual images. In the emerging mutimodal dialogue systems, this information becomes essential for performance improvement. We thus take a step further towards understanding and adaptively using such multimodal evidence.

2.2 Multimodal Understanding

Another line of work related to ours lies in multimodal understanding, which focuses on recognizing inherent semantics of multimodal data [1] and exploits the relevance between different modalities [2, 7, 27]. In this work, we focus on understanding concept-level semantics within multimodal data, which relates to research of visual-semantic embedding, image captioning and visual question-answering (VQA).

With the aim of learning a mapping from images into a semantic space, visual-semantic embedding have shown to be effective in image-text ranking and zero-shot learning. There are some embedding models based on Canonical Correlation Analysis (CCA) [11] which learns a linear projection to maximize the correlation between two modalities [8, 9]. Kernel CCA [22] is further employed to extend to nolinear projection. Nevertheless, as point out in [38], scaling CCA to large amounts of data can be difficult. Another line of efforts train a joint embedding model with ranking loss. [7] learns linear transformation between visual and textual features with a single-directional ranking loss, which applies penalty to incorrect sentences ranked higher than correct ones. Bi-direction ranking loss is employed to boost the performance by further ensuring the correct image described by a sentence ranked higher than other images [19–21]. However, these works cannot be directly utilized by our work due to the difference in nature between our dialogue state tracking task and other tasks. In the above tasks, the semantics of image and text are to be aligned, while in the multimodal dialogue problem, the semantics of user intent come from either visual supplied evidence or textual inputs. Though certain semantics are presented by both modality, many of them only reside in one modality due to the concise nature of dialogues.

For image captioning, most existing methods adopt an encoder-decoder framework, which consists of a CNN visual encoder along with an RNN language decoder [27, 37]. The CNN encoder extracts visual feature from the image and feed it to the RNN decoder to generate a natural language text. Inspired by the advances in neural machine translation, some works further introduced the attention mechanism, which attends to different sub-regions of an image during decoding [42].

Compared to image captioning, VQA [2] shares a more similar task setting with multimodal dialogue, since it involves single-turn interaction through question and answer. VQA also utilized the encoder-decoder framework while the encoder side includes both image and question [2, 32]. The work of visual dialogue [4, 29] further handles multi-turn QA pairs as multimodal dialogue does. Even so, as pointed out in [29], these works should be categorized

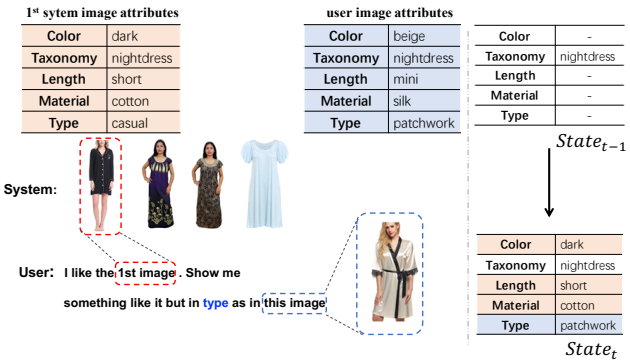

Figure 2: An example turn in multimodal dialogue. The model extracts semantics from both visual and textual input to update the dialogue state to $State_t$. Slots "Color", "Length" and "Material" are updated by understanding the system provided image. While "Type" is updated based on both textual and visual evidence from the user provided image.

into image-grounded QA rather than multimodal dialogue. The reason lays on the way they utilize images. In VQA tasks, only one single image is involved and the conversation is centered on it. While in multimodal dialogue tasks, the model has to process multiple images which acts as supporting evidence.

However, as pointed out in [10, 45], most existing Vision-to-Language task performance should be attributed to the language prior and the models do not truly understand the images. While in our scenario, accurate understanding of visual image is rather important for intention inference. Therefore, we propose to extract explicit semantic concepts from image and feed to downstream dialogue tracker.

Recently, [33] contributed a large-scale benchmark dataset with 150K dialogue sessions and proposed two end-to-end models based on the hierarchical recurrent encoder decoder (HRED) framework [35] as baselines for response generation. We use their dataset while focus on the multimodal dialogue state tracking task.

3 MODEL

3.1 Task Definition and Model Overview

Our model is designed to tackle the problem of dialogue state tracking (DST) which maintains the belief state $\mathbf{S}_t$ in each turn during the dialogue flow. Different from traditional text-based task, we need to extract semantic concepts from both textual and visual modalities. An example is shown in Figure 2. Note that the slot values of a state are derived from not only the textual evidence, but also system provided and user provided visual evidences.

Suppose $T_t^{'}, V_t^{'}$ refer to the textual and visual parts of the system response in turn t, while T_t, V_t are those of the subsequent user posts, the task of multimodal dialogue state tracking can be stated as follows: in the t-th turn, given the dialogue context with messages from both user and agent sides $X_t = \{T_1^{'}, V_1^{'}, T_1, V_1, ..., T_t^{'}, V_t^{'}, T_t, V_t\}$, the model is supposed to give an estimation of the belief state $\mathbf{p}_t^k$ for each slot k:

$$\mathbf{p}_t^k = f^k(X_t), \tag{1}$$

where f^k represents the belief tracker for slot k, $\mathbf{p}_t^k$ is a $N_k + 1$ dimensional vector[1] which is a probability distribution over the values of slot k. The belief state $\mathbf{S}_t$ in turn t is defined as a collection of distributions over the values of all the N_s slots:

$$\mathbf{S}_t = [\mathbf{p}_t^1, \mathbf{p}_t^2, ..., \mathbf{p}_t^{N_s}]. \tag{2}$$

As described above, both the system and user messages may consist of a textual and a visual part where the textual part is a natural language utterance and the visual part refers to some images. It is worth noting that in our work, the visual part of system response $V_t^{'}$ usually contains zero or multiple images, while the user post V_t contains zero or only one image.

Specifically, our proposed model consists of three major parts:

- The basic textual network learns representations of the textual evidences by taking only the textual utterances from both system and user sides as inputs:

$$\mathbf{h}_t = f_{textual}(T_1^{'}, T_1, T_2^{'}, T_2, ..., T_t^{'}, T_t). \tag{3}$$

- The sub-region based visual concept learning part extracts concept level visual representation $\mathbf{v}$ for each image I. Given the image, it first embeds sub-regions and then maps to concepts under the multiple instance learning scenario:

$$\mathbf{v} = f_{visual}(I). \tag{4}$$

- The adaptive modality attention part predicts the belief state by integrating the textual and visual representations:

$$\mathbf{p}_t = f_{attn}(\mathbf{v}_t^{'}, \mathbf{v}_t, \mathbf{r}_t, \mathbf{h}_t), \tag{5}$$

where $\mathbf{v}_t^{'}$ and $\mathbf{v}_t$ denote the visual representation of system and user images respectively, $\mathbf{r}_t$ refers to the textual representation of user text T_t. The model learns to automatically emphasize on different sources of evidences for dialogue state tracking.

In the following part, we will explain the three parts in more detail. Since all the slots share a common model structure, for the convinience of description, we omit the slot subscript k in the following subsections and only describe the model architecture for a single slot.

3.2 Basic Textual Framework

We first introduce a basic textual framework upon which we will build our multimodal model. The framework is an hierarchical RNN model, which consists of a word-level encoder and an utterance-level encoder. In this framework, the textual inputs in the t-th turn are T_t and $T_t^{'}$. Suppose $\mathbf{T}_t$ and $\mathbf{T}_t^{'}$ are the embedding matrices obtained via one hot vectors of words in user post T_t and system response $T_t^{'}$ multiplying with the pre-trained word embedding matrix respectively. Under such processing, we actually represent each word in its semantic vector form, namely word vectors. We then employ a word-level RNN encoder, which takes as input a word vector at each time step, to learn the integrated semantic representations of T_t and $T_t^{'}$ respectively as follows:

$$\mathbf{r}_t = RNN_1(\mathbf{T}_t), \tag{6}$$

$$\mathbf{r}_t^{'} = RNN_1(\mathbf{T}_t^{'}), \tag{7}$$

where the word-level encoder is denoted as RNN_1. The concatenation of $\mathbf{r}_t$ and $\mathbf{r}_t^{'}$ is then fed into an utterance-level encoder, which is denoted as RNN_2. Therefore, we have

$$\mathbf{h}_t = RNN_2\_CELL(\mathbf{h}_{t-1}, [\mathbf{r}_t, \mathbf{r}_t^{'}]), \tag{8}$$

where $\mathbf{h_t}$ is the output of RNN cell, and [,] denotes concatenation of vectors.

Note that $\mathbf{h_t}$ is the summary of information observed so far in turn t. In pure text-based systems, they usually directly fed $\mathbf{h_t}$ into a fully-connected layer followed by a softmax function to generate the probability distribution over the values of the slot:

$$\mathbf{g}_t = softmax(FC(\mathbf{h}_t)). \tag{9}$$

The dimension of the fully-connected layer is $d + 1$, where d is the number of possible slot values for this specific slot. The extra one dimension represents that this slot is not mentioned yet. As mentioned before, we have a totally of N_s slots. Thus, we maintain N_s such RNN pipelines for all the slots.

3.3 Sub-region based Visual Concept Learning

One of the major challenges in multimodal DST is to accurately extract visual concepts from images, which can be easily formulated as a multi-class or multi-label image classification problem with manual annotation. However, we notice that most visual concepts correspond to only a small sub-region of the image, such as the slots "Neck" and "Belt-loops". Studies like [6] have also shown that feeding the whole image into attribute classifiers leads to worse performance. Therefore, we delve into image sub-regions to harvest visual concepts.

Basically, for each individual image i, we define a bag b_i, which is a collection of the image's sub-regions (detailed in Section 4.3). An image is labeled positive for the concept w if there is at least one sub-region in the bag containing w, which is actually a multiple instance learning (MIL) problem. In this work, the sub-regions are squared areas which can overlap in case that some objects are cut up by the square boundary. Intuitively, suppose the probability of one sub-region j containing the concept w is v_{ij}^w, the probability of the whole image i containing w should be no less than the largest sub-region concept probability in the bag. At the same time, multiple sub-regions showing high probability of containing the concept w should result in increased probability but not over-exaggerated. A diminishing return characteristic is preferable for our case. Therefore, the probability of image i containing the concept w is defined as follows:

$$v_i^w = 1 - \prod_{j=1}^{|b_i|}(1 - v_{ij}^w), \tag{10}$$

where v_i^w is not larger than 1 and not less than the largest v_{ij}^w. When there are more v_{ij}^w being large, the incremental effect on v_i^w actually decreases.

We use the ResNet-50 without the last layer as the base network to learn the representation for image sub-region b_{ij} as $\mathbf{b}_{ij}$, and feed

[1] N_k is the number of values of slot k, while the extra one dimension represents that slot k is not mentioned in the dialogue yet.

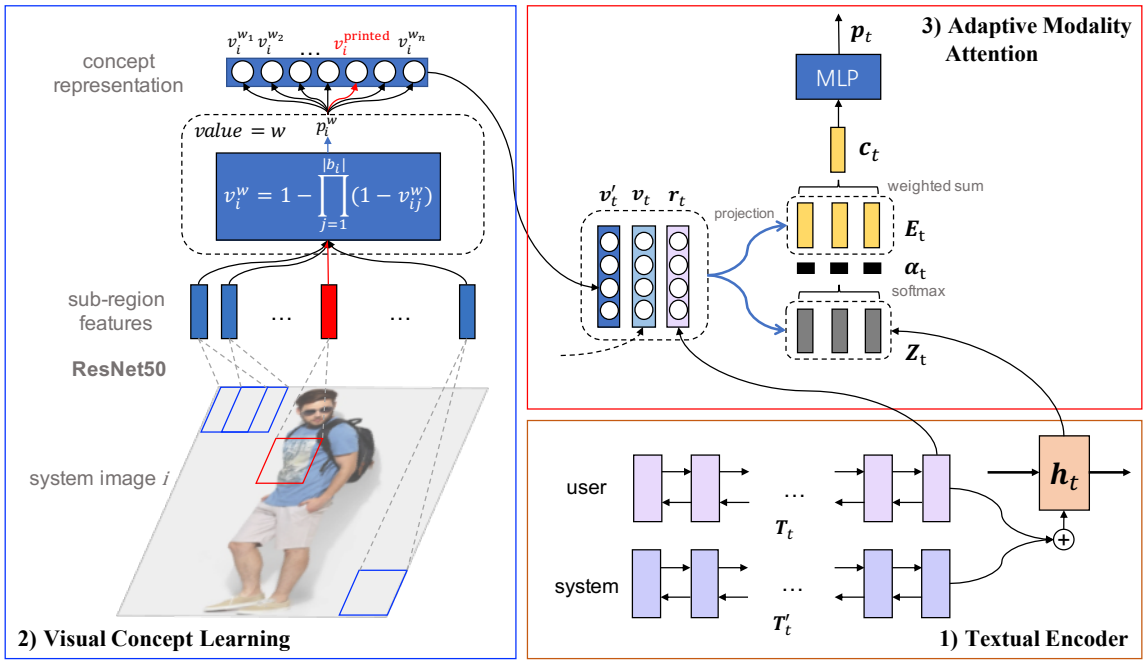

Figure 3: Overview of our proposed model. We first extract the features of user's and system's textual input evidences using two RNN networks and use them to update the textual base hidden state $\mathbf{h}_t$. In the visual concept learning stage, the model learns visual semantics by looking into image sub-regions and integrating probabilities under the multiple instance learning scenario. Finally, we apply the adaptive modality attention mechanism on different evidence sources to obtain the context vector $\mathbf{c}_t$, which is then fed into a MLP followed by a softmax activation to get the final belief state $\mathbf{P}_t$.

it into a fully-connected layer followed by a sigmoid activation to compute the sub-region level concept probability vector $\mathbf{v}_{ij}$:

$$\mathbf{v}_{ij} = \sigma(FC(\mathbf{b}_{ij}))), \tag{11}$$

where $\mathbf{v}_{ij}$ is a collection of Bernoulli distributions over each concept and v_{ij}^w is the probability of concept w.

At the whole image level, we obtain the image representation $\mathbf{v}_i$ for image i via

$$\mathbf{v}_i = 1 - \prod_{j=1}^{|b_i|}(1 - \mathbf{v}_{ij}). \tag{12}$$

3.4 Adaptive Modality Attention

When updating the dialogue state, each slot should emphasize on different evidence sources of the input. For example in Figure 2, the user said "... something like it but in type as in this image", which means that the value of "type" should be equal to that of the user given image, while the values of other primary slots should be the same as the chosen image ("the 1st image"). Another common situation is that we can find an exact value in user text, such as "Show me more in purple colored type".

Generally speaking, there are three main sources of evidences for tracking dialogue states: the system provided images, user provided image [2] and user provided text. Note that the system provided text offers little useful clues about states, we thus only use it in context modeling as in Equation 8, and do not take it as an evidence source in a way similar to [30].

Essentially, we aim to get a context vector $\mathbf{c}_t$ as an attentive summarization of the three evidence sources: user provided image representation $\mathbf{v}_t$, system provided image representation $\mathbf{v}_t^{'}$, and user text representation $\mathbf{r}_t$. We then feed it into a fully-connected layer followed by a softmax activation to find the new distribution on values:

$$\mathbf{p}_t = softmax(FC(\mathbf{c}_t)). \tag{13}$$

The summarization vector $\mathbf{c}_t$ is obtained by combining the representations of the three evidence sources using a weight vector $\boldsymbol{\alpha}_t$, which is an probability distribution over the three input sources. How to decide the attention weights $\boldsymbol{\alpha}_t$ is essential to our model since emphasizing on the correct evidence source is critical for generating the right slot value. Therefore, we first use two projection matrices to map the three evidence vectors into a common space and obtain an concatenated evidence matrix $\mathbf{E}_t$:

$$\mathbf{c}_t = \sum_{m=1}^{3} \alpha_{t,m}\mathbf{E}_t(m), \tag{14}$$

$$\mathbf{E}_t = [\mathbf{W}_1\mathbf{v}_t, \mathbf{W}_1\mathbf{v}_t^{'}, \mathbf{W}_2\mathbf{r}_t], \tag{15}$$

where $\mathbf{W}_1 \in \mathcal{R}^{h\times d}$ and $\mathbf{W}_2 \in \mathcal{R}^{h\times h}$ are projection parameters. m indicates the index of evidence source and $\mathbf{E}_t(m)$ is the m-th column vector of $\mathbf{E}_t$.

To get the attention distribution $\boldsymbol{\alpha}_t$, we first project the visual and textual evidences into a common space and then feed them

[2] In the dataset, users only provide one image each time.

along with the RNN output $\mathbf{h}_t$ into a neural network followed by a softmax activation:

$$\mathbf{Z}_t = [\mathbf{W}_3\mathbf{v}_t, \mathbf{W}_3\mathbf{v}_t^{'}, \mathbf{W}_4\mathbf{r}_t], \tag{16}$$

$$k_{t,m} = score(\mathbf{Z}_t(m), \mathbf{h}_t), \tag{17}$$

$$\alpha_{t,m} = \frac{\exp(k_{t,m})}{\sum_{i=1}^{3}\exp(k_{t,m})}, \tag{18}$$

where $\mathbf{W}_3 \in \mathcal{R}^{h\times d}$, $\mathbf{W}_4 \in \mathcal{R}^{h\times b}$ are projection parameters mapping the textual and visual evidence $\mathbf{v}_t$, $\mathbf{v}_t^{'}$ and $\mathbf{r}_t$ into a common space. The *score* function yields a scalar measuring to what extent each evidence source is matched to the slot. It is based on the projected evidence source $\mathbf{Z}_t(m)$ and the textual encoder state $\mathbf{h}_t$. In our implementation, we parametrized it as a feed forward neural network which is jointly trained with all the other components.

$$score(\mathbf{Z}_t(m), \mathbf{h}_t) = \mathbf{w}_h^{\top}\tanh\left(\mathbf{Z}_t(m) + \mathbf{W}_g\mathbf{h}_t\right), \tag{19}$$

where $\mathbf{w}_h \in \mathcal{R}^h$, $\mathbf{W}_g \in \mathcal{R}^{h\times h}$ are the parameters to be learned, h, b are the model hyper-parameters in which b is the RNN hidden state size. Note that in our attention mechanism, we use different key and value matrices $\mathbf{Z}_t$ and $\mathbf{E}_t$ by projecting the evidence source into different spaces, since earlier works [28] have already suggested that the dual use of a single vector makes training the model difficult.

The intuition of this design is as follows: as defined in Section 3.2, the utterance-level hidden state $\mathbf{h}_t$ of the basic textual framework contains the long and short term information obtained from the textual side. More specifically, the long term information is an aggregation of previous dialogue evidences, such as whether certain slot has already been mentioned in history. While the short term information includes more recent evidences such as which slot is mentioned in the current user post. Both kinds of information can provide extensive clues on which modality should be attended to in the current turn. We thus feed the evidence source representations along with $\mathbf{h}_t$ through a single layer neural network (the *score* function) followed by a softmax activation to generate the attention distribution $\boldsymbol{\alpha}_t$ over different evidence sources.

We use cross entropy loss to measure the prediction results of belief tracker. Specifically, we have:

$$L = -\sum_{w} y_t^w \log p_t^w, \tag{20}$$

where w indicates a certain slot value, p_t^w is an element of $\mathbf{p}_t$ which is the probability that w is chosen as the new belief state, and y_t^w is the golden truth.

4 EXPERIMENT

In this section, we conducted extensive experiments to validate our NMBT's performance on the task of dialogue state tracking[3]. More specifically, we want to figure out:

- How much can the task of dialogue state tracking benefit by involving multimodal evidence ?
- Whether the image representations obtained by sub-region based visual concept learning perform better than the dense features extracted by the CNN models ?
- Has the adaptive attention mechanism really learned to pay attention to the correct modality ?

[3]The code is available at https://github.com/zhangzthu/NMBT

4.1 Data Preparation

The collection of training corpus is one of the bottlenecks for developing statistical dialogue system, due to the cost and concerns about privacy disclosure. Recently, [34] proposed a large-scale multimodal dialogue dataset which consists of about 150K conversation sessions between sale agents and shoppers. However, there is no dialogue state labels in the original dataset, we thus utilized the meta data provided by [34] to build the domain ontology[4] for dialogue act annotation. More specifically, we extracted slot-value pairs from the corpus using a two stage process. During the first stage, we defined a slot value dictionary and 81 natural language templates to extract slots using direct string matching techniques. During the second stage, we conducted manual correction. Finally, seven slots are considered and the number of their corresponding values are shown in Table 1. To train the sub-region based visual concept learning model, we also need to get a collection of images with labels in this domain. We extracted the labels of each image from the raw catalog which consists of text descriptions of each fashion item.

Table 1: The number of values of different slots.

Slot Name	# Value
material	114
style	17
color	47
type	56
fit	17
length	19
gender	3

In the dataset, the average turn number of dialogue sessions is 18.3, which is a lot larger than other text-based datasets [13, 14] and leads to greater difficulty.

4.2 Baselines

To evaluate the effectiveness of our proposed NMBT model, we compared it with the following baselines.

- **Seq2seq_DST** [17]: This model includes an encoder-decoder architecture with an attention mechanism to map an input utterance to a sequence of slot-value pairs. Note that Seq2seq_DST is different from RNN_DST in that the last hidden state of Seq2seq_DST's encoder is fed into an RNN decoder, rather than a MLP classifier.
- **CNN_DST**: A CNN based textual neural belief tracker[36], which utilizes a slot-specific filter to extract semantic features from raw inputs.
- **RNN_DST**: The textual only framework of our NMBT model. The hidden state of the utterance level encoder is fed into a MLP classifier to predict the belief state.
- **NBT**: A textual DST model which uses CNN to learn n-gram utterance feature from word embeddings. The utterance feature is then used for DST tracking [30].
- **M-RNN_DST**: We extended the RNN_DST model to multimodal scenario by feeding the image feature vectors of each turn as extra input of the untterance level encoder, which is similar to

[4]The domain ontology means all the slots and their values to be involved in the study.

the Multimodal HRED model in [34]. The difference is that the last hidden state of the untterance level encoder is fed into a MLP with a softmax activation to predict the slot value rather than a natural language decoder.

- **M-NBT**: The multimodal version of NBT, in which we combined the ResNet-50 extracted image features with the text features obtained before the last layer of NBT to predict the dialogue state, similar as Multimodal HRED [34] does.

It is worth noting that Seq2seq_DST, CNN_DST, RNN_DST and NBT are representative textual-based models from the Dialogue State Tracking Challenge [13], while M-RNN_DST and M-NBT are two multimodal extensions based on strong textual baselines.

Table 2: The dialogue state tracking accuracy of different models.

Method	Overall	Slots			
		Style	Material	Color	Fit
Seq2seq_DST	51.1	80.0	84.7	47.7	51.2
CNN_DST	51.7	81.4	81.5	50.8	54.8
RNN_DST	51.2	80.4	85.1	49.5	51.5
NBT	52.0	84.9	66.1	49.3	47.9
M-RNN_DST	56.4	81.8	87.0	49.9	53.7
M-NBT	57.2	82.7	87.7	51.5	56.1
NMBT w/o SBVL	58.6	86.9	89.7	52.6	59.0
NMBT w/o Attn	57.5	81.3	88.0	54.3	56.8
NMBT	**59.8**	**87.9**	**92.3**	**55.6**	**60.6**

4.3 Experimental Setups

The training of our model is carried out in two stages. First, we pre-train the basic textual model and the visual concept detector respectively. Then we fuse them together to train the full NMBT model with adaptive modality attention. For the textual base component, both RNN encoders' hidden state sizes are set to 512, and the number of layers is 3. We use 300 as the dimension of word vectors, which are extracted by pre-trained GloVe model [31]. During training, we keep the extracted word embeddings fixed.

The sub-regions in visual concept learning are defined by sliding window. More specifically, we first resize the image to 468×468, and the size of image sub-region is 224×224 as in ResNet-50. With a stride size of 30, we finally get 81 (9×9) sub-regions for each image. The stride size is determined through mode validation, since a smaller size leads to computing complexity and a larger one may cut off a potential object. The size of image feature extracted by ResNet-50 is 2048, which is then fed into a MLP followed by a sigmoid activation to get $\mathbf{v}_{ij}$ (see Equation 10). The h in Section 3.4 is set to 100. Both MIL and adaptive modality attention parameters is trained using Adam optimizer with a learning rate of 0.001, and the momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4.4 Evaluation Protocols

We adopted the dialogue state tracking accuracy as the main metric to evaluate the performance of our model. In each turn, the model predicts a value for each slot k. For each slot, we can get a slot-wise accuracy. The overall accuracy is averaged over all slots. Each user utterance from the original corpus is annotated with a specific *state type*, which indicates its function, such as "*show orientations*" and "*show similar to*". To give a more elaborate comparison, we also provide the accuracy scores of each baseline model on different state types.

We also analysis the performance of visual concept detection. For each slot k, we picked out the value w with the highest probability v_{ij}^w as the predicted value for that slot. In this metric, we show the overall and slot-wise accuracy scores. To give a more intuitional understanding of how the visual concept extractor works, we visualize its v_{ij}^w results for some concept w in Figure 4. As described in Section 4.3, we get a 9×9 v_{ij}^w map for each image. In order to visualize these probabilities on the image, we simply resize the heat map to 464×464 by upsampling and apply a Gaussian filter. Note that the heat map here is not the "attention" heat map which is widely used in many attention-based visual models [42]. In our model, v_{ij}^w is a binary classification probability, which indicates the probability that sub-region j contains concept w.

To validate the effectiveness of our adaptive modality attention mechanism, we conduct a case study with a visualization of the adaptive attention weights. Due to space limitation, we only visualize the attention weights of five representative slots.

4.5 Performance of Dialogue State Tracking

We first report the DST prediction performance on the overall accuracy score and accuracy scores of several representative slots. Besides the baseline methods described in Section 4.2, we also conduct two ablation studies on the sub-region based visual concept learning (**SBVL**) component and modality attention mechanism (**Attn**). The results are shown in Table 2. Note that we only report the result of some primary slots which occupy about xx% of the data.

Some key observations are summarized as follows:

- First of all, as compared to textual-based methods, there is a significant improvement on both the overall and the slot-specific accuracy scores for those methods considering multimodal information. For example, the overall accuracy score of M-RNN_DST increases 5.2% as compared to that of its pure textual-based version RNN_DST. The overall accuracy score of M-NBT also shows similar improvement pattern comparing to that of NBT. It indicates that by involving images into dialogue state tracking, the model is able to extract more useful information for better tracking the user's intention. In fact, it is natural to use images in the fashion products shopping conversation scenario. As the example shown in Figure 1, the images carry detailed information about the user's requirements which can not be easily expressed using only textual utterances. Therefore, it is crucial to capture the visual evidence in dialogue state tracking under these multimodal scenarios.
- Secondly, by learning sub-region based concept level visual representations, our proposed model achieves better performance as compared to the multimodal models that leverage pre-trained model extracted visual features. We compared with several multimodal baselines and a variation of our model named NMBT w/o SBVL, in which we remove the sub-region based visual concept

Table 3: The comparison of DST accuracy of different methods over three state type, where "*show similar to*" and "*like show result*" indicate the user asks to recommend products similar to the current one, and "*like earlier show result*" means the user requires products similar to someone in previous turns. Several representative slot-specific accuracy scores and the average accuracy score are reported. Note that the accuracy scores here are calculated on state type level, which can not be compared with the results in Table 2.

Method	*show similar to*					*like show result*					*like earlier show result*				
	Style	**Material**	**Color**	**Fit**	**Overall**	**Style**	**Material**	**Color**	**Fit**	**Overall**	**Style**	**Material**	**Color**	**Fit**	**Overall**
Seq2seq_DST	81.5	88.6	49.4	52.1	52.6	86.2	88.0	47.9	55.7	53.6	81.7	83.6	48.6	46.2	51.6
CNN_DST	84.3	85.7	51.7	56.4	53.7	87.6	85.8	51.1	60.3	54.7	83.8	80.7	51.5	51.0	52.5
RNN_DST	83.3	87.4	49.9	55.3	52.6	88.0	87.3	48.0	55.1	54.0	81.9	86.0	48.7	46.6	52.4
NBT	84.7	72.5	45.3	50.7	52.0	85.5	71.9	54.6	47.3	53.1	84.0	67.6	49.0	45.2	52.1
M-RNN_DST	81.2	82.3	50.0	67.1	64.9	78.2	88.2	47.3	71.8	66.9	**87.5**	82.9	51.0	49.0	60.7
M-NBT	84.3	84.6	50.5	67.4	65.7	85.0	91.2	52.1	71.6	69.0	83.0	84.9	51.4	46.8	60.1
NMBT	**87.6**	**91.1**	**55.4**	**72.1**	**68.5**	**88.6**	**95.9**	**55.7**	**75.5**	**71.2**	85.6	**88.9**	**55.0**	**51.6**	**61.8**

learning component while replace it with the pre-trained ResNet-50 to extract visual features. In Table 2, we observe that utilizing visual information by simply using representations learned by pre-trained ResNet-50 is not an efficient way. There exists a large performance gap between our method NMBT and M-RNN_DST, M-NBT. For instance, the overall accuracy score of NMBT is improved by 3.4% and 2.6% respectively as compare to that of M-RNN_DST and M-NBT. For our proposed model NMBT, when the sub-region based visual concept learning component is removed, the overall performance of resulting method NMBT w/o SBVL drops about 1.2%. These results validate the effectiveness of our proposed visual concept learning model. By using sub-regions of image rather than the whole image, the model manages to learn more accurate concepts with less background noise.

- Thirdly, by combining textual and visual evidences through adaptive modality attention mechanism, NMBT manages to learn a better integrated representation of multimodal evidences. In M-RNN_DST and M-NBT, the textual and visual representations are integrated by vector concatenation. In the ablation model NMBT w/o Attn, we removed the attention mechanism by redefining the context vector in Equation 13 as a concatenation of the column vectors in $\mathbf{E}_t$. Compared to these multimodal models, NMBT achieves better performance on both slot-specific and overall accuracy. Intuitively, it is a non-trivial task to fuse the visual and textual evidences together. The image provided by either system or user can be regarded as an attribute list, which acts as a candidate value set for belief state update. The textual utterances contain two important clues: the values of some slots which are explicitly expressed in the utterance, such as "looking for some **formal** shoes", and the control information which decides the source of each slot's update, such as "in the **style** as in the **1st** image". This mechanism can not be model by either simple feature concatenation as in M-RNN_DST, M-NBT, or the NMBT w/o Attn ablation which simply averages the three evidences. In contrast, our model learns to automatically emphasize on the different evidence sources for updating belief state based on the conversational context.

4.6 Fine-grained DST Performance Analysis

To give a more detailed analysis, we report the state type level accuracy on three representative types, as shown in Table 3. First we analyze the function of "*show similar to*" which means that the user requires the agent to show similar items to the currently selected product. The selected product can come from one of the three evidence sources, which favors our adaptive attention mechanism. Therefore, we observe that the performance of our method NMBT in this type is better than the averaged performance shown in Table 2, which means in this scenario where the modality choice appears frequently, the attention mechanism manages to emphasize the correct modality in most cases. Also, our NMBT method obtains higher accuracy compared to the baseline methods. Next we analyzed "*like show result*" which is similar to "*show similar to*" and thus we can get similar conclusion based on its results, which further verifies NMBT's performance on emphasizing different modalities. When coming to the "*like earlier show result*" type which means that the user requires items related to results shown several turns before, we observed that although NMBT still outperforms baseline methods, the margin is dwindled from 2.2% to 1.1%. The possible reason might be that in "*like earlier show result*", the agent needs to recall back to previous turns other than the current turn, which brings extra difficulty.

4.7 Evaluation of Visual Concept Learning Component

Learning good visual representations is crucial to the model's performance since it serves as the input to downstream modules of dialogue state tracking. We thus analyze the effectiveness of visual concept learning in terms of concept classification accuracy and report results in Table 4.

Table 4: The visual concept detection accuracy on different slots.

Method	Overall	Slots			
		Style	**Material**	**Color**	**Fit**
ResNet50	42.8	58.7	21.4	**42.8**	48.1
SRVL	**44.8**	**61.2**	**24.2**	42.0	**52.3**

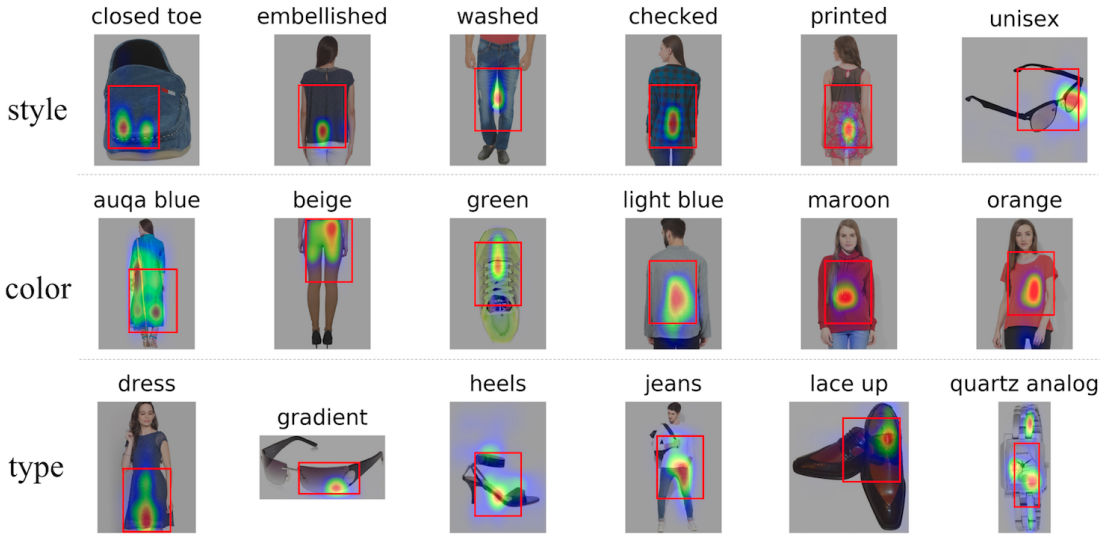

Figure 4: Sub-region based multiple instance learning for visual concept learning. We visualized the concept region probability of values such as *style*, *color* and *type*, where the three slots correspond to the three rows. The sub-region with the highest probability containing the concept is marked by a red rectangle.

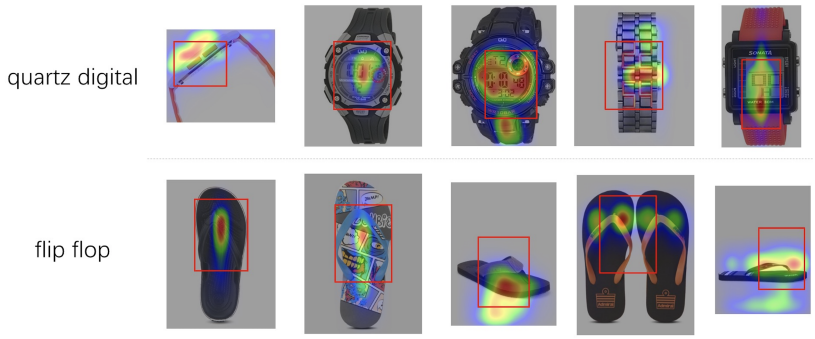

Figure 5: Visualization of the heat map of v_{ij}^{w} and the bounding box of *quartz digital* and *flip flop* concepts. We present multiple examples for the two concepts where the product images are taken in different orientations.

We found that the sub-region based visual concept learning (**SRVL**) can capture more accurate classification features by localizing the visual concept and reducing background noise. This statement can be concluded from Table 4, where the overall and several representative slot-wise concept prediction accuracy scores are reported. We compare with the ResNet-50 model which directly uses the feature vector of the entire image to predict a concept. As shown in Table 4, by incorporating sub-region based visual concept learning, the accuracy score is improved by about 2.0%. Nevertheless, as it can be seen in the slot-specific result, the accuracy on "*color*" is dwindled by about 0.8%. This result is generally in line with our intuition, since the slot "*color*" is a more holistic concept which makes it more easily to be predicted with the overall image feature.

In order to validate whether our design of using image sub-regions is reasonable and whether our model can find those correct sub-regions for concepts, we visualize the spatial response map v_{ij}^{w} of some concepts as shown in Figure 4. The bounding box of the sub-region with the highest probability containing the concept is marked by a red rectangle. We can see from the visualization results that without bounding box annotations for training, our model is still able to locate and associate visual concepts with correct sub-regions. For example, in the "*closed toe*" picture, our model locates the sub-region on the toe part of a shoe, and in the "*washed*" example, our model focuses on the washed-styled pants rather than the T-shirts. For "*color*" prediction, the model focuses on the main body of each product which takes the most information about the product's color. These results indicate that the image representations given by sub-region based multiple instance learning indeed capture important visual classification concepts and can thus extract informative features for the task.

We further validate whether our proposed model can capture a concept presented in different orientations or poses.In fashion

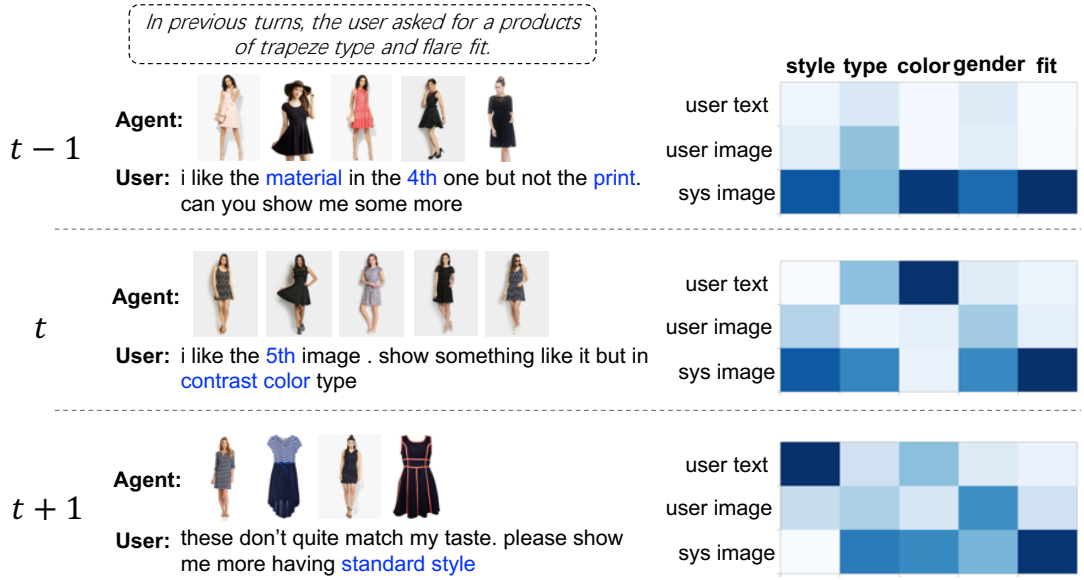

Figure 6: Visualization of modality attention weights. The attention weights of five major slots over the three evidence sources are plotted, where darker color indicates larger attention weight. Note that each column represents an attention distribution.

domain, the product images are often taken in different orientations, which requires the visual concept detector to identify a single concept in different modes. For example, as for the concept "*watch*", the model needs to identify it no matter its image is taken from the side or front, while the image features can be very different in this two cases. To validate whether our model fits this requirement, we conducted a more specific analysis by showing the heat map and bounding box of a single concept in images taken from different orientations. As demonstrated in Figure 5, we gave five images for two concepts, taken from at least three different orientations. For the concept of "*quartz digital*", our model can identify it by locating the feature of *knob*, *dial* and *band*, which are very different features. As for "*flip flop*", our model identifies its key characteristic, the *toe thong*, from different orientations. In the 4th image where there are two of it, the model is even able to find the *toe thong* feature for both identical objectives.

4.8 Case Study of Adaptive Modality Attention

As the adaptive modality attention is essential in our model, we conducted case studies by visualizing the modality attention weights to verify how it works in modeling multimodal conversations. As shown in Figure 6, we visualized the attention weights for our proposed model. Note that each column represents an attention distribution over the three sources where darker color indicates larger attention weight.

The result demonstrates that the adaptive modality attention mechanism is able to automatically emphasize the three sources based on both textual and visual contexts in multimodal dialogue, even without specific supervision. For example in Figure 6, during the t-th turn, the user asked for products similar to the 5th one but with a different color, which indicates that the update of all slots except "*color*" should attend to system image. The corresponding attention weights fit this expectation, in which the colors of other slots is more shaded on "*sys image*", while that of "*color*" is more shaded on user text. In the $(t+1)$-th turn, the user asked for "*standard*" style products which is explicitly expressed in user text. Hence, the tracker of "*style*" pays more attention to user text. This case shows that the adaptive attention mechanism can capture the clues about which part is more important for each slot's update and thus boosts the overall performance.

5 CONCLUSION

In this work, we studied how visual evidence can be incorporated in the task of dialogue state tracking, and proposed a neural multimodal belief tracking model named NMBT, which seamlessly integrates and adaptively selects textual and visual information in multi-model dialogues. This model consists of a textual encoder which encodes textual utterances, a sub-region based visual concept detector which extracts concepts from image, and a multi-modality attention mechanism which adaptively attends to textual or visual evidence during conversations. Extensive experiments demonstrated that our model outperforms the state-of-the-art baselines. Results showed that dialogue state tracking in multimodal dialogues can significantly benefit from jointly considering multi-modal evidences.

Multi-modal dialogue systems have many applications in real-world dialogue systems such as online shopping or virtual dialogue agent. We believe that this research direction is still in its infancy and our work may inspire many future studies.

6 ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (Grant No. 2018YFC0830200), the National Science Foundation of China (Grant No.61876096/61332007) and the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Xavier Alameda-Pineda, Miriam Redi, Mohammad Soleymani, Nicu Sebe, Shih-Fu Chang, and Samuel Gosling. 2017. MUSA2: First ACM Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1974–1975.

[2] 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.

[3] Dan Bohus and Alex Rudnicky. 2006. A k-hypotheses+ other belief updating model. In *Proc. of the AAAI Workshop on Statistical and Empirical Methods in Spoken Dialogue Systems*, Vol. 62.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.

[5] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2 (1997), 31–71.

[6] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*. 1473–1482.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[8] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106, 2 (2014), 210–233.

[9] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. Springer, 529–545.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*. 3.

[11] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16, 12 (2004), 2639–2664.

[12] Matthew Henderson. 2015. Machine learning for dialog state tracking: A review. In *Proc. of The First International Workshop on Machine Learning in Spoken Language Processing*.

[13] Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 263–272.

[14] Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The third dialog state tracking challenge. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 324–329.

[15] Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*. 467–471.

[16] Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 292–299.

[17] Takaaki Hori, Hai Wang, Chiori Hori, Shinji Watanabe, Bret Harsham, Jonathan Le Roux, John R Hershey, Yusuke Koji, Yi Jing, Zhaocheng Zhu, et al. 2016. Dialog state tracking with attention-based sequence-to-sequence learning. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 552–558.

[18] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239.

[19] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[20] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*. 1889–1897.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539* (2014).

[22] Pei Ling Lai and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* 10, 05 (2000), 365–377.

[23] Staffan Larsson and David R Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering* 6, 3-4 (2000), 323–340.

[24] Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1998. Using Markov decision process for learning dialogue strategies. In *ICAssP*, Vol. 98. 201–204.

[25] Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing* 8, 1 (2000), 11–23.

[26] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 801–809.

[27] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014).

[28] Hideya Mino, Masao Utiyama, Eiichiro Sumita, and Takenobu Tokunaga. 2017. Key-value Attention Mechanism for Neural Machine Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 290–295.

[29] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251* (2017).

[30] Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777* (2016).

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[32] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*. 2953–2961.

[33] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2017. Multimodal Dialogs (MMD): A large-scale dataset for studying multimodal domain-aware conversations. *arXiv preprint arXiv:1704.00200* (2017).

[34] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2017. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. *arXiv preprint arXiv:1704.00200* (2017).

[35] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*, Vol. 16. 3776–3784.

[36] Hongjie Shi, Takashi Ushio, Mitsuru Endo, Katsuyoshi Yamagami, and Noriaki Horii. 2017. Convolutional neural networks for multi-topic dialog state tracking. In *Dialogues with Social Robots*. Springer, 451–463.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[38] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[39] Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse* 7, 3 (2016), 4–33.

[40] Jason D Williams. 2012. A critical analysis of two statistical spoken dialog systems in public use. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*. 55–60.

[41] Jason D Williams. 2014. Web-style ranking and SLU combination for dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 282–291.

[42] 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.

[43] Steve Young. 2006. Using POMDPs for dialog management. In *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, 8–13.

[44] Steve J Young. 2000. Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 358, 1769 (2000), 1389–1402.

[45] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*. 5014–5022.

[46] Victor Zue, Stephanie Seneff, James R Glass, Joseph Polifroni, Christine Pao, Timothy J Hazen, and Lee Hetherington. 2000. JUPITER: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing* 8, 1 (2000), 85–96.