Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems

Zheng Zhang\textsuperscript{1,}\textsuperscript{*}, Lizi Liao\textsuperscript{2,}\textsuperscript{†}, Minlie Huang\textsuperscript{1,}\textsuperscript{†}, Xiaoyan Zhu\textsuperscript{1}, Tat-Seng Chua\textsuperscript{2}

\textsuperscript{1}Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems \\
\textsuperscript{2}Beijing National Research Center for Information Science and Technology \\
\textsuperscript{3}Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China \\
\textsuperscript{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

• Computing methodologies → Discourse, dialogue and pragmatics; • Information systems → Information retrieval; Multimedia and multimodal retrieval; Users and interactive retrieval.

ACM Reference Format:

1 INTRODUCTION

User: I am looking for some formal shoes with liberty type patterns
System: Sorry I don’t have but would you like to see some as below

User: Show me something similar to the 4th image
System: The similar looking ones are

User: What type is it in the 1st image?
System: The formal shoe in the 1st image has formal type
User: Show me more in the style as in the 1st image
System: Found some as

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 straightforward 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

\textsuperscript{*} Corresponding author.
\textsuperscript{†} These two authors contributed equally.
The main contributions for this work are as follows:

1. 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.
2. We delve into image sub-regions to learn concept level semantics and propose an adaptive attention mechanism for automatically deciding the evidence source for dialogue state tracking based on multimodal dialogue context.
3. 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 approaches 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 multimodal 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 nonlinear 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 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 \( S_t \) 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 image. While “Type” is updated based on both textual and visual evidence from the user provided image.

![Figure 2: An example turn in multimodal dialogue. The model extracts semantics from both visual and textual input to update the dialogue state to \( S_t \). 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.](image-url)
We first introduce a basic textual framework upon which we will apply a word-level RNN encoder, which takes as input a word vector $T_t$ of distributions over the values of slot $k$. The belief state $S_t$ in turn $t$ is defined as a collection of distributions over the values of all the $N_k$ slots:

$$S_t = [p^1_t, p^2_t, \ldots, p^{N_k}_t].$$

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 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:

$$h_t = f_{\text{textual}}(T_t^1, T_t^2, \ldots, T_t^1, T_t).$$

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

$$v = f_{\text{visual}}(I).$$

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

$$p_t = f_{\text{attn}}(v_t, v_t, r_t, h_t),$$

where $v_t'$ and $v_t$ denote the visual representation of system and user images respectively, $r_t$ 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 convenience 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 $T_t$ and $T'_t$ 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:

$$r_t = RNN_1(T_t),$$

$$\hat{r}_t = RNN_2(T'_t).$$

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

$$h_t = RNN_{2,\text{CELL}}(h_{t-1}, [r_t, \hat{r}_t]).$$

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 $N_k$ slots. Thus, we maintain $N_k$ 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:

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

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 $h_{ij}$, and feed

---

$N_k$ is the number of values of slot $k$, while the extra one dimension represents that slot $k$ is not mentioned in the dialogue yet.
it into a fully-connected layer followed by a sigmoid activation to compute the sub-region level concept probability vector $v_{ij}$:

$$v_{ij} = \sigma(FC(b_{ij})), \quad (11)$$

where $v_{ij}$ is a collection of Bernoulli distributions over each concept and $w_i$ is the probability of concept $w$.

At the whole image level, we obtain the image representation $v_i$ for image $i$ via

$$v_i = 1 - \prod_{j=1}^{[b_i]} (1 - v_{ij}). \quad (12)$$

### 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 $c_t$ as an attentive summarization of the three evidence sources: user provided image representation $v_t$, system provided image representation $v_t'$, and user text representation $r_t$. We then feed it into a fully-connected layer followed by a softmax activation to find the new distribution on values:

$$p_t = \text{softmax}(FC(c_t)). \quad (13)$$

The summarization vector $c_t$ is obtained by combining the representations of the three evidence sources using a weight vector $\alpha_t$, which is a probability distribution over the three inputs. How to decide the attention weights $\alpha_t$ 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 $E_t$:

$$c_t = \sum_{m=1}^{3} \alpha_{t,m} E_t(m), \quad (14)$$

$$E_t = [W_1 v_t, W_1 v_t', W_2 r_t], \quad (15)$$

where $W_1 \in \mathbb{R}^{h \times d}$ and $W_2 \in \mathbb{R}^{h \times h}$ are projection parameters. $m$ indicates the index of evidence source and $E_t(m)$ is the $m$-th column vector of $E_t$.

To get the attention distribution $\alpha_t$, we first project the visual and textual evidences into a common space and then feed them
along with the RNN output $h_t$ into a neural network followed by a softmax activation:

$$Z_t = [W_3 v_t, W_3 v_t, W_4 h_t],$$

$$k_{t,m} = \text{score}(Z_t(m), h_t),$$

$$a_{t,m} = \frac{\exp(k_{t,m})}{\sum_{j=1}^{3} \exp(k_{t,j})},$$

where $W_3 \in \mathbb{R}^{h \times d}$, $W_4 \in \mathbb{R}^{h \times b}$ are projection parameters mapping the textual and visual evidence $v_t$, $v_t$, and $h_t$ into a common space. The $\text{score}$ function yields a scalar measuring to what extent each evidence source is matched to the slot. It is based on the projected evidence source $Z_t(m)$ and the textual encoder state $h_t$. In our implementation, we parameterized it as a feed forward neural network which is jointly trained with all the other components.

$$\text{score}(Z_t(m), h_t) = w_h^\top \tanh(Z_t(m) + W_4 h_t),$$

where $w_h \in \mathbb{R}^h$, $W_4 \in \mathbb{R}^{h \times h}$ are the parameters to be learned, $h$, $b$ are the model hyper-parameters in which $h$ is the RNN hidden state size. Note that in our attention mechanism, we use different key and value matrices $Z_t$ and $E_t$ 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 $h_{t}$ 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 $h_{t}$ through a single layer neural network (the $\text{score}$ function) followed by a softmax activation to generate the attention distribution $a_t$ over different evidence sources.

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

$$L = - \sum w_t \log p_t^w,$$

where $w$ indicates a certain slot value, $p_t^w$ is an element of $p_t$ 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>
<thead>
<tr>
<th>Slot Name</th>
<th># Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>material</td>
<td>114</td>
</tr>
<tr>
<td>style</td>
<td>17</td>
</tr>
<tr>
<td>color</td>
<td>47</td>
</tr>
<tr>
<td>type</td>
<td>56</td>
</tr>
<tr>
<td>fit</td>
<td>17</td>
</tr>
<tr>
<td>length</td>
<td>19</td>
</tr>
<tr>
<td>gender</td>
<td>3</td>
</tr>
</tbody>
</table>

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 utterance 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 utterance 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>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Style</td>
</tr>
<tr>
<td>Seq2Seq_DST</td>
<td>51.1</td>
<td>80.0</td>
</tr>
<tr>
<td>CNN_DST</td>
<td>51.7</td>
<td>81.4</td>
</tr>
<tr>
<td>RNN_DST</td>
<td>51.2</td>
<td>81.4</td>
</tr>
<tr>
<td>M-RNN_DST</td>
<td>56.4</td>
<td>81.8</td>
</tr>
<tr>
<td>NBT</td>
<td>52.0</td>
<td>84.9</td>
</tr>
<tr>
<td>M-NBT</td>
<td>57.2</td>
<td>82.7</td>
</tr>
<tr>
<td>NMBT w/o Attn</td>
<td>58.6</td>
<td>86.9</td>
</tr>
<tr>
<td>M-NBT</td>
<td>57.5</td>
<td>81.3</td>
</tr>
<tr>
<td>NMBT</td>
<td>59.8</td>
<td>87.9</td>
</tr>
</tbody>
</table>

**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 \(v_{ij}\) (see Equation 10). The \(\beta\) in Section 3.4 is set to 100. Both MLP 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 analyze 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 intuitive 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 \(9 \times 9 v_{ij}^{w}\) map for each image. In order to visualize these probabilities on the image, we simply resize the heat map to \(464 \times 464\) 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...
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 E_v. 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 analyze “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 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>show similar to</th>
<th>like show result</th>
<th>like earlier show result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Style</td>
<td>Material</td>
<td>Color</td>
</tr>
<tr>
<td>Seq2seq_DST</td>
<td>81.5</td>
<td>86.6</td>
<td>49.4</td>
</tr>
<tr>
<td>CNN_DST</td>
<td>84.3</td>
<td>85.7</td>
<td>51.7</td>
</tr>
<tr>
<td>RNN_DST</td>
<td>83.3</td>
<td>87.4</td>
<td>49.9</td>
</tr>
<tr>
<td>NBT</td>
<td>84.7</td>
<td>72.5</td>
<td>45.3</td>
</tr>
<tr>
<td>M-RNN_DST</td>
<td>81.2</td>
<td>82.3</td>
<td>50.0</td>
</tr>
<tr>
<td>M-NBT</td>
<td>84.3</td>
<td>84.6</td>
<td>50.5</td>
</tr>
<tr>
<td>NMBT</td>
<td>87.6</td>
<td>91.1</td>
<td>55.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Style</td>
</tr>
<tr>
<td>ResNet50</td>
<td>42.8</td>
<td>58.7</td>
</tr>
<tr>
<td>SRVL</td>
<td>44.8</td>
<td>61.2</td>
</tr>
</tbody>
</table>
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
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.

### 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-modal 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 multimodality 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 multimodal 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.