Knowledge Aware Emotion Recognition in Textual Conversations via Multi-Task Incremental Transformer

Duzhen Zhang[1,2], Xiuyi Chen[1,2*], Shuang Xu[1] and Bo Xu[1,2]

[1]Institute of Automation, Chinese Academy of Sciences (CASIA). Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

{zhangduzhen2019,chenxiuyi2017,shuang.xu,xubo}@ia.ac.cn

Abstract

Emotion recognition in textual conversations (ERTC) plays an important role in a wide range of applications, such as opinion mining, recommender systems, and so on. ERTC, however, is a challenging task. For one thing, speakers often rely on the context and commonsense knowledge to express emotions; for another, most utterances contain neutral emotion in conversations, as a result, the confusion between a few non-neutral utterances and much more neutral ones restrains the emotion recognition performance. In this paper, we propose a novel Knowledge Aware Incremental Transformer with Multi-task Learning (KAITML) to address these challenges. Firstly, we devise a dual-level graph attention mechanism to leverage commonsense knowledge, which augments the semantic information of the utterance. Then we apply the Incremental Transformer to encode multi-turn contextual utterances. Moreover, we are the first to introduce multi-task learning to alleviate the aforementioned confusion and thus further improve the emotion recognition performance. Extensive experimental results show that our KAITML model outperforms the state-of-the-art models across five benchmark datasets.

1 Introduction

Emotion recognition in textual conversations (ERTC), which aims to identify the emotion of each utterance from the transcript of a conversation, has become a popular research topic in recent years. ERTC can be widely used in various scenarios, such as opinion mining of comments in social media (Chatterjee et al., 2019), emotion analysis of customers in artificial customer service, and others. In addition, it can also be applied to chat robots to analyze the user's emotional state in real time and generate emotion-aware responses (Poria et al., 2019b; Zhou et al., 2018a; Huang et al., 2018).

Truth	Prediction			
	Others	Angry	Sad	Happy
Others	4424	**101**	**60**	**92**
Angry	**54**	237	6	1
Sad	**44**	11	192	3
Happy	**88**	0	2	194

Table 1: A confusion matrix of the emotion recognition results on the EmoContext test dataset (Chatterjee et al., 2019) from Knowledge-Enriched Transformer (Zhong et al., 2019b), which is the current state-of-the-art model. We notice that there are barely miss-classifications among the non-neutral categories (Angry, Sad, and Happy). Most of the errors, shown in the bold font, correspond to the confusion between a few non-neutral categories and much more neutral category (Others).

However, there are several challenges when analyzing emotion in natural conversations. Firstly, unlike vanilla emotion recognition of sentences (Wang and Manning, 2012; Seyeditabari et al., 2018), ERTC requires comprehensively considering the context in the conversation. Secondly, knowledge plays an

*The first two authors contributed equally.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4429–4440
Barcelona, Spain (Online), December 8-13, 2020

important role in ERTC as speakers often express emotions relying on the context and commonsense knowledge (Zhong et al., 2019b). Moreover, most utterances contain neutral emotion in conversations, and the heavily imbalanced class distribution can easily lead to the confusion between a few non-neutral utterances (e.g., happy, sad, and angry, etc.) and much more neutral ones (e.g., neutral or others), which restrains the emotion recognition performance. Table 1 shows a confusion matrix of emotion recognition results from the current state-of-the-art model and there appears a serious confusion between a few non-neutral categories and much more neutral category.

Some prior studies have been conducted to model contextual information for emotion recognition in conversations (Poria et al., 2017; Majumder et al., 2019). These methods first adopt convolutional neural networks (CNN) to extract utterance-level features and then use context-level recurrent neural networks (RNN) to model the contextual utterances in conversation. However, RNN and CNN have difficulty modeling long-distance dependencies (Vaswani et al., 2017), which may be useful in ERTC. Zhong et al. (2019b) uses a context-aware affective graph attention mechanism to incorporate external knowledge for ERTC. However, they don't consider various relations in external knowledge base, which may cause the loss of semantic information. In addition, to the best of our knowledge, no existing work considers the confusion between a few non-neutral utterances and much more neutral ones.

In this paper, We propose a novel Knowledge Aware Incremental Transformer with Multi-task Learning (KAITML) to address the aforementioned challenges. Firstly, we enhance the background and semantic information of the given utterance to facilitate ERTC with the retrieved relevant knowledge graphs from a large-scale commonsense knowledge base. Specifically, we propose a dual level graph attention mechanism to encode these relevant knowledge graphs, which consists of a node-level attention to learn the importance of different neighboring nodes and a relation-level attention to learn the importance of different relations to the current node. Then we apply the Incremental Transformer (Li et al., 2019) to incrementally encode multi-turn contextual utterances, which could capture the intra-utterance and inter-utterance correlations by the self-attention (Cheng et al., 2016) and context-attention (Zhang et al., 2018) modules, respectively. Moreover, we introduce multi-task learning to alleviate the confusion between a few non-neutral utterances and much more neutral ones, as shown in Table 1. Specifically, we first focus on the binary classification, "non-neutral" versus "neutral", and then classifies the "non-neutral" ones into fine-grained emotion categories. These two auxiliary tasks are jointly trained with the original emotion recognition task.

In summary, this paper makes the following contributions:

- We devise a dual-level graph attention mechanism to support better understanding of utterances for ERTC by considering various relations in external knowledge base. Furthermore, we apply the Incremental Transformer to model multi-turn contextual utterances and recognize emotions.
- We are the first to introduce multi-task learning with two auxiliary tasks to alleviate the aforementioned confusion and thus further improve emotion recognition performance.
- Experimental results show that our proposed KAITML model outperforms the state-of-the-art models across five benchmark datasets in F1 score. In addition, context, commonsense knowledge and multi-task learning are all beneficial to the emotion recognition performance.

2 Related Work

Emotion recognition in conversations has grabbed much attention from researchers in the past few years due to the proliferation of publicly available conversational dataset (Poria et al., 2019a; Chatterjee et al., 2019; Li et al., 2017; Zhou et al., 2018a) and its widespread applications in opinion mining, recommender systems, emotion-aware dialogues generation, and so on (Poria et al., 2019b). Some of the deep learning-based models have been proposed for emotion recognition in conversations, in only textual and multimodal settings (containing textual, acoustic, and visual information).

Poria et al. (2017) proposes a long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) based model to capture contextual correlations from the utterances of a user-generated video

for multimodal sentiment classification. Hazarika et al. (2018b) proposes conversational memory network (CMN) that exploits distinct memory units for each speaker to model emotional dynamics and detect emotion in a dyadic conversation. Later, Hazarika et al. (2018a) improves upon this approach with interactive conversational memory network (ICON), which utilizes the interactive memory unit to hierarchically model the self- and inter-speaker emotional influences for emotion recognition in conversational videos. Majumder et al. (2019) proposes the DialogueRNN model that exploits three gated recurrent units (GRU) (Cho et al., 2014) to capture speaker information, context and emotional information of the preceding utterances, respectively. They achieve the state-of-the-art performance on several multimodal conversation datasets. Compared to these gated RNNs and CNNs based models, we apply the Incremental Transformer (Li et al., 2019) to incrementally encode multi-turn contextual utterances, where the shorter path of information flow in the self-attention (Cheng et al., 2016) and context-attention (Zhang et al., 2018) modules in the Incremental Transformer allows our model to exploit contextual information more efficiently.

Recently, a considerable literature has grown up around the theme of incorporating external knowledge in generative conversation systems, including question answering systems (Hao et al., 2017; Mihaylov and Frank, 2018), open-domain dialogue systems (Young et al., 2018; Zhou et al., 2018b; Zhong et al., 2019a), and task-oriented dialogue systems (He et al., 2019; Madotto et al., 2018; Chen et al., 2019). Zhong et al. (2019b) proposes a Knowledge-Enriched Transformer (KET) achieving the state-of-the-art performance on multiple textual conversation datasets, where contextual utterances are encoded using hierarchical self-attention and commonsense knowledge is incorporated using a context-aware affective graph attention mechanism. However, they ignore various relations in external knowledge base, which may cause the loss of semantic information. By contrast, our dual-level graph attention mechanism, can take advantage of the various relations in external knowledge base to better augment the semantic information of the utterances.

3 Our Proposed KAITML Model

3.1 Task Definition and Overview

Let $\langle X_j^{(i)}, Y_j^{(i)} \rangle, i = 1, ...N, j = 1, ...N_i$ be a collection of $\langle utterance, label \rangle$ pairs in a given conversation dataset, where N denotes the number of conversations and N_i denotes the number of utterances in the ith conversation. The objective of the task is to maximize the following function:

$$\arg\max_{\theta} \prod_{i=1}^{N} \prod_{j=1}^{N_i} \mathrm{P}(Y_j^{(i)} \mid X_j^{(i)}, X_{j-1}^{(i)}, ..., X_{j-M}^{(i)}; \theta). \tag{1}$$

where $X_j^{(i)}$ denotes target utterance, $Y_j^{(i)}$ denotes the emotion label of target utterance, θ denotes the model parameters we need to optimize and $X_{j-1}^{(i)}, ..., X_{j-M}^{(i)}$ denote contextual utterances. Here, we limit the number of contextual utterances to M. We follow (Su et al., 2018; Zhong et al., 2019b) to directly discard early contextual utterances. Similar to (Zhong et al., 2019b; Poria et al., 2017), we clip and pad each utterance $X_j^{(i)}$ to a fixed K number of tokens. The overview of our KAITML model and detailed architecture of model components are presented in Figure 1.

3.2 Knowledge Interpreter

Commonsense knowledge is fundamental to understanding conversations (Zhou et al., 2018b). We use ConceptNet (Speer et al., 2017) as a external commonsense knowledge base in our model. ConceptNet is a large-scale multilingual semantic graph where concepts are nodes in the graph and relations are edges, which describes general human knowledge in natural language. Each $\langle concept1, relation, concept2 \rangle$ triple is termed an assertion. At present, ConceptNet comprises 5.9M assertions, 3.1M concepts and 38 relations for English.

The knowledge interpreter is designed to facilitate the understanding of an utterance. It takes as input an utterance $X_n^{(i)} = x_1 x_2 ... x_K, n = j - M, ..., j$ and retrieves a few relevant knowledge graphs $G_n^{(i)} =$

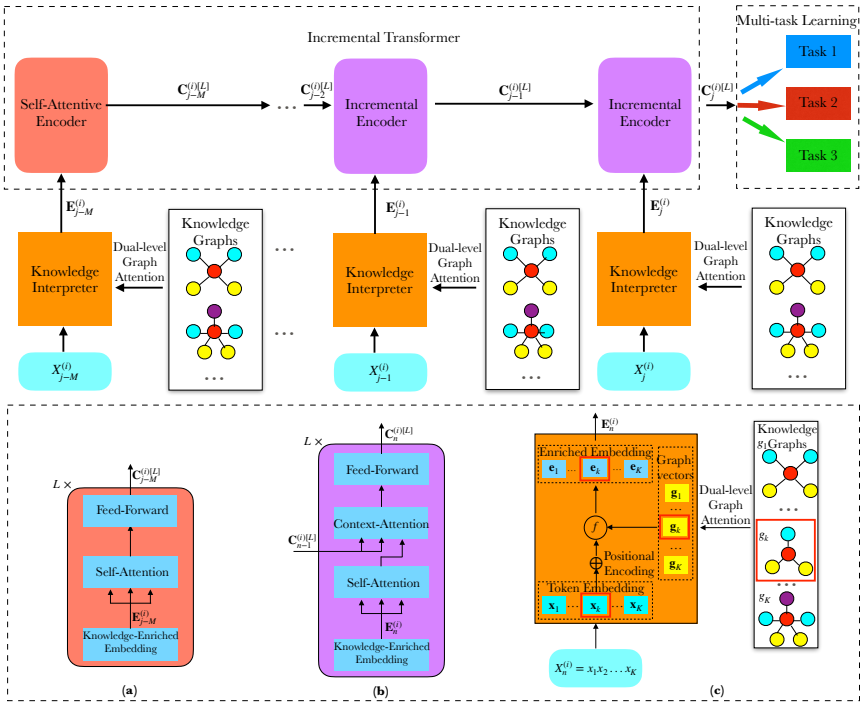

Figure 1: The top is the overview of KAITML and the bottom is the detailed architecture of model components. (a) Self-Attentive Encoder. (b) Incremental Encoder. (c) Knowledge Interpreter.

$\{g_1, g_2, ..., g_K\}$ where each token in the utterance corresponds to a graph, as shown in Figure 1 (c). In general, the knowledge interpreter uses each token $x_k, k = 1, ..., K$ (non-stopword) in an utterance $X_n^{(i)}$ as the key node to retrieve a graph g_k comprising its immediate neighbors from ConceptNet, as shown in the red box in Figure 1 (c). For each g_k, we remove nodes that are stopwords or not in our vocabulary. Each retrieved graph g_k consists of a key node (the red dots) and its neighboring nodes (different colors denote different relations), where each node c is converted into a vector representation $\boldsymbol{c} \in \mathbb{R}^d$, where d denotes the size of vector. Then, the knowledge interpreter computes the graph vector $\boldsymbol{g}_k \in \mathbb{R}^d$ of the retrieved graph g_k using the dual-level graph attention mechanism.

We use a token embedding layer to convert each token x_k in $X_n^{(i)}$ into a vector representation $\boldsymbol{x}_k \in \mathbb{R}^d$. To encode positional information, the position encoding (Vaswani et al., 2017) is added as follows:

$$\boldsymbol{x}_k = \text{Embed}(x_k) + \text{Pos}(x_k). \tag{2}$$

Finally, the knowledge-enriched token embedding $\boldsymbol{e_k}$ can be obtained via a linear transformation:

$$\boldsymbol{e}_k = \boldsymbol{W}[\boldsymbol{x}_k; \boldsymbol{g}_k]. \tag{3}$$

where $[;]$ denotes concatenation and $\boldsymbol{W} \in \mathbb{R}^{d \times 2d}$ denotes a model parameter. All K tokens in $X_n^{(i)}$ form a knowledge-enriched utterance embedding $\boldsymbol{E}_n^{(i)} \in \mathbb{R}^{K \times d}$ that is then fed to the Incremental Transformer, as shown in Figure 1.

Dual-level Graph Attention Mechanism

The dual-level graph attention mechanism is designed to generate a representation for a retrieved knowledge graph, inspired by (Velickovic et al., 2018), which will be used to augment the semantics of each token in an utterance. Compared to (Velickovic et al., 2018), our graph attention considers not only all nodes in a graph but also relations between nodes. The dual-level graph attention mechanism, including node-level and relation-level attentions, can learn the importance of different neighboring nodes as well as the importance of different relations to a key node.

Node-level Attention. The node-level attention takes as input the node vectors $\boldsymbol{F}(g_k) = \{\boldsymbol{c}_s^r\}, r = 1, ...R_k, s = 1, ...N_r$ in the retrieved knowledge graph g_k, where R_k denotes the number of relations in g_k and N_r denotes the number of nodes in the rth relation, to produce relation vectors $\boldsymbol{t}_r$ as follows:

$$\boldsymbol{t}_r = \sum_{s=1}^{N_r} \alpha_s^r \boldsymbol{c}_s^r, \tag{4}$$

$$\alpha_s^r = \frac{\exp(\boldsymbol{x}_k \cdot \boldsymbol{c}_s^r)}{\sum_{h=1}^{N_r} \exp(\boldsymbol{x}_k \cdot \boldsymbol{c}_h^r)}. \tag{5}$$

Relation-level Attention. The relation-level attention takes as input the relation vectors $\boldsymbol{t}_r, r = 1, ...R_k$, to produce a graph vector $\boldsymbol{g}_k$ as follows:

$$\boldsymbol{g}_k = \sum_{r=1}^{R_k} \beta_r \boldsymbol{t}_r, \tag{6}$$

$$\beta_r = \frac{\exp(\boldsymbol{x}_k \cdot \boldsymbol{t}_r)}{\sum_{h=1}^{R_k} \exp(\boldsymbol{x}_k \cdot \boldsymbol{t}_h)}. \tag{7}$$

If $|g_k| = 0$, where $|g_k|$ denotes the number of nodes in g_k, we set $\boldsymbol{g}_k$ to the average of all node vectors (Zhong et al., 2019b).

3.3 Incremental Transformer

We apply the Incremental Transformer (Li et al., 2019) to encode multi-turn contextual utterances, as shown in Figure 1, which contains Self-Attentive Encoder and Incremental Encoder.

Self-Attentive Encoder

The Self-Attentive Encoder is a transformer encoder as described in (Vaswani et al., 2017), which encodes the first utterance.

As shown in Figure 1 (a), the Self-Attentive Encoder contains a stack of L identical layers. Each layer has two sub-layers. The first sub-layer is a multi-head self-attention (MultiHead) (Vaswani et al., 2017). $MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ is a multi-head attention function that takes a query matrix $\boldsymbol{Q}$, a key matrix $\boldsymbol{K}$, and a value matrix $\boldsymbol{V}$ as input. In current case, $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V}$. That's why it's called self-attention. And the second sub-layer is a simple, position-wise fully connected feed-forward network (FFN). This FFN consists of two linear transformations with a ReLU activation in between, $FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$, where W_1, b_1, W_2, b_2 denote model parameters. (Vaswani et al., 2017)

Formally, for the first knowledge-enriched utterance embedding $\boldsymbol{E}_{j-M}^{(i)} \in \mathbb{R}^{K \times d}$, its representation $\boldsymbol{C}_{j-M}^{(i)[L]} \in \mathbb{R}^{K \times d}$ is computed as follows:

$$\boldsymbol{A}_{j-M}^{(i)[l]} = \text{MultiHead}(\boldsymbol{C}_{j-M}^{(i)[l-1]}, \boldsymbol{C}_{j-M}^{(i)[l-1]}, \boldsymbol{C}_{j-M}^{(i)[l-1]}), \tag{8}$$

$$\boldsymbol{C}_{j-M}^{(i)[l]} = \text{FFN}(\boldsymbol{A}_{j-M}^{(i)[l]}). \tag{9}$$

where $l = 1, ..., L$, $\boldsymbol{C}_{j-M}^{(i)[0]} = \boldsymbol{E}_{j-M}^{(i)}$, $\boldsymbol{A}_{j-M}^{(i)[l]} \in \mathbb{R}^{K \times d}$ is the hidden state computed by multi-head attention at the lth layer, $\boldsymbol{C}_{j-M}^{(i)[l]} \in \mathbb{R}^{K \times d}$ denotes the representation of $\boldsymbol{E}_{j-M}^{(i)}$ after l layer. The residual connection and layer normalization are omitted in the presentation for simplicity. More details can be found in (Vaswani et al., 2017).

Incremental Encoder

The Incremental Encoder is a variant of the transformer encoder with an additional context-attention (Zhang et al., 2018) module, which encodes multi-turn utterances using an incremental encoding scheme. It takes the output of previous utterances and current utterance as input, and use attention mechanism to incrementally model relevant context.

As shown in Figure 1 (b), the Incremental Encoder contains a stack of L identical layers. Each layer has three sub-layers. For each knowledge-enriched utterance embedding $\boldsymbol{E}_n^{(i)} \in \mathbb{R}^{K\times d}, n = j - M + 1, ..., j$, its representation $\boldsymbol{C}_n^{(i)[L]} \in \mathbb{R}^{K\times d}$ is computed as follows:

The first sub-layer is a multi-head self-attention:

$$\boldsymbol{A}_n^{(i)[l]} = \text{MultiHead}(\boldsymbol{C}_n^{(i)[l-1]}, \boldsymbol{C}_n^{(i)[l-1]}, \boldsymbol{C}_n^{(i)[l-1]}), \tag{10}$$

where $l = 1, ...L$, $\boldsymbol{C}_n^{(i)[l-1]} \in \mathbb{R}^{K\times d}$ is the output of the previous layer and $\boldsymbol{C}_n^{(i)[0]} = \boldsymbol{E}_n^{(i)}$.

The second sub-layer is a multi-head context-attention:

$$\boldsymbol{B}_n^{(i)[l]} = \text{MultiHead}(\boldsymbol{A}_n^{(i)[l]}, \boldsymbol{C}_{n-1}^{(i)[L]}, \boldsymbol{C}_{n-1}^{(i)[L]}), \tag{11}$$

where $\boldsymbol{C}_{n-1}^{(i)[L]} \in \mathbb{R}^{K\times d}$ is the representation of the previous utterances after L layers.

The third sub-layer is a position-wise fully connected feed-forward network:

$$\boldsymbol{C}_n^{(i)[l]} = \text{FFN}(\boldsymbol{B}_n^{(i)[l]}). \tag{12}$$

Finally, $\boldsymbol{C}_j^{(i)[L]} \in \mathbb{R}^{K\times d}$ is the representation of relevant context (including target utterance), as shown in Figure 1, which is then fed into a max-pooing layer to learn discriminative features among positions and derive the final representation $\boldsymbol{O}_j^{(i)} \in \mathbb{R}^d$:

$$\boldsymbol{O}_j^{(i)} = \text{MaxPooling}(\boldsymbol{C}_j^{(i)[L]}). \tag{13}$$

3.4 Multi-task Learning

We introduce multi-task learning to alleviate the confusion between a few non-neutral categories (e.g., happy, sad, and angry, etc.) and much more neutral category (e.g., neutral or others) and thus further improve emotion recognition performance, as shown in Figure 1, which contains three different tasks.

Task 1 is the original emotion recognition task, which predicts the emotion label, including non-neutral categories and neutral category, of target utterance $X_j^{(i)}$. Its loss on one sample $\langle X_j^{(i)}, Y_j^{(i)} \rangle$ is computed as follows:

$$loss_1 = -\sum_{t=1}^{q} {Y_1}_{j_t}^{(i)} \log \hat{Y_1}_{j_t}^{(i)}, \tag{14}$$

$$\hat{\boldsymbol{Y}_{\boldsymbol{1}}}_j^{(i)} = \text{softmax}(\boldsymbol{O}_j^{(i)} W_1 + b_1). \tag{15}$$

where $W_1 \in \mathbb{R}^{d\times q}$ and $b_1 \in \mathbb{R}^q$ denotes model parameters, q denotes the number of categories, $\hat{\boldsymbol{Y}_{\boldsymbol{1}}}_j^{(i)} \in \mathbb{R}^q$ denotes the predicted probability distribution of task 1, and $\boldsymbol{Y_1}_j^{(i)} \in \mathbb{R}^q$ (one-hot vector, the corresponding category position is 1, and the remaining positions are 0) denotes the ground-truth probability distribution of task 1.

Task 2 focuses on the binary classification, "non-neutral" versus "neutral", which determines whether the target utterance $X_j^{(i)}$ is "non-neutral" or "neutral". Its loss on one sample $\langle X_j^{(i)}, Y_j^{(i)} \rangle$ is computed as follows:

$$loss_2 = -\sum_{t=1}^{2} {Y_2}_{j_t}^{(i)} \log \hat{Y_2}_{j_t}^{(i)}, \tag{16}$$

$$\hat{\boldsymbol{Y}_{\boldsymbol{2}}}_j^{(i)} = \text{softmax}(\boldsymbol{O}_j^{(i)} W_2 + b_2). \tag{17}$$

where $W_2 \in \mathbb{R}^{d\times 2}$ and $b_2 \in \mathbb{R}^2$ denotes model parameters, $\hat{\boldsymbol{Y}}_{\boldsymbol{2}}{}_j^{(i)} \in \mathbb{R}^2$ denotes the predicted probability distribution of task 2, and $\boldsymbol{Y}_{\boldsymbol{2}}{}_j^{(i)} \in \mathbb{R}^2$ (one-hot vector) denotes the ground-truth probability distribution of task 2.

Task 3 classifies the "non-neutral" into fine-grained emotion categories. Its loss on one sample $\langle X_j^{(i)}, Y_j^{(i)} \rangle$ is computed as follows:

$$loss_3 = \frac{1}{2}\sum_{t=1}^{q-1}(Y_{3_{j_t}}^{(i)} - \hat{Y}_{3_{j_t}}^{(i)})^2, \tag{18}$$

$$\hat{\boldsymbol{Y}}_{\boldsymbol{3}}{}_j^{(i)} = \text{sigmoid}(\boldsymbol{O}_j^{(i)} W_3 + b_3). \tag{19}$$

where $W_3 \in \mathbb{R}^{d\times(q-1)}$ and $b_3 \in \mathbb{R}^{q-1}$ denotes model parameters, $q-1$ denotes the number of non-neutral categories, $\hat{\boldsymbol{Y}}_{\boldsymbol{3}}{}_j^{(i)} \in \mathbb{R}^{q-1}$ denotes the predicted output of task 3, and $\boldsymbol{Y}_{\boldsymbol{3}}{}_j^{(i)} \in \mathbb{R}^{q-1}$ denotes the ground-truth output of task 3 (when $Y_j^{(i)}$ is neutral category, $\boldsymbol{Y}_{\boldsymbol{3}}{}_j^{(i)}$ is a vector of all zeros, otherwise it's a one-hot vector).

During training, the total loss of our model is defined as:

$$Loss = \frac{loss_1 + \lambda_1 loss_2 + \lambda_2 loss_3}{1 + \lambda_1 + \lambda_2}. \tag{20}$$

where $\lambda_1, \lambda_2 \in [0, 1]$ are weight coefficients of $loss_2, loss_3$, respectively.

4 Experimental Setting

4.1 Datasets and Evaluations

We evaluate our model on the following five benchmark datasets. Some of datasets, such as MELD, IEMOCAP, EmoryNLP, are multimodal conversation datasets containing textual, acoustic, and visual information. In this paper, we recognize emotion in conversations only based on textual information. The statistics and evaluation metrics of these datasets are drawn in Table 2.

Dataset	#Conv. (Train/Val/Test)	#Utter. (Train/Val/Test)	#Classes	Evaluation
EC	30160/2755/5509	90480/8265/16527	4	Micro-F1(exclude "others")
DailyDialogue	11118/1000/1000	87170/8069/7740	7	Micro-F1(exclude "neutral")
MELD	1038/114/280	9989/1109/2610	7	Weighted-F1
EmoryNLP	659/89/79	7551/954/984	7	Weighted-F1
IEMOCAP	100/20/31	4810/1000/1523	6	Weighted-F1

Table 2: Dataset descriptions.

EmoContext(EC) (Chatterjee et al., 2019): Short dialogues composed of three turns comes from social media. Its emotion labels include happy, sad, angry and others.

DailyDialogue (Li et al., 2017): Daily communications written by human. Its emotion labels include anger, disgust, fear, joy, sadness, surprise and neutral.

MELD (Poria et al., 2019a): Scripts collected from the Friends TV series. Its emotion labels are the same as the ones used in DailyDialogue.

EmoryNLP (Zahiri and Choi, 2018): Scripts collected from the Friends TV series as well. Its emotion labels include sad, mad, scared, powerful, peaceful, joyful and neutral, which are different from MELD.

IEMOCAP (Busso et al., 2008): Two-way emotional conversation. Its emotion labels include happiness, sadness, anger, frustrated, excited and neutral.

The evaluation metric of each dataset is the same as the one used in (Zhong et al., 2019b).

4.2 Baselines

We compare our proposed model with the following baselines:
cLSTM: It first adopt a bidirectional LSTM to extract utterance-level features and then use a context-level unibidirectional LSTM to model the contextual utterances.
CNN (Kim, 2014): A single-layer CNN is trained on utterance-level without context.
CNN+cLSTM (Poria et al., 2017): It first adopt an CNN to extract utterance-level features and then apply a context-level unibidirectional LSTM to learn context representations.
BERT_BASE (Devlin et al., 2019): Base version of Bert. It takes as input each utterance with its context as a single text.
DialogueRNN (Majumder et al., 2019): It exploits three gated recurrent units (GRU) to capture speaker information, context and emotional information of the preceding utterances, respectively.
KET (Zhong et al., 2019b): It's the state-of-the-art model for ERTC, where contextual utterances are encoded using hierarchical self-attention and commonsense knowledge is incorporated using a context-aware affective graph attention mechanism.

4.3 Hyper-parameter Settings

We use Adam optimizer (Kingma and Ba, 2015) to train our model with learning rate of 0.0001 and a batch size of 64 in all datasets. We set the class weights in cross-entropy loss as the ratio of the class distribution in the validation set to the class distribution in the training set for each dataset (Zhong et al., 2019b). Thus, we can tackle the mismatch in class distribution between validation set and training set. The initial token and node embeddings are pre-trained with GloVe (Pennington et al., 2014). The detailed hyper-parameter settings for KAITML are presented in Table 3.

Dataset	d	p	f	M	L	h	λ_1	λ_2
EC	300	400	1	2	2	4	1.0	0.7
DailyDialogue	300	400	3	6	3	4	1.0	1.0
MELD	300	400	1	6	1	4	1.0	0.7
EmoryNLP	100	200	1	6	2	4	0.9	0.5
IEMOCAP	300	400	1	6	1	4	0.9	0.6

Table 3: Hyper-parameter settings for KAITML. d: token/node embedding size. p: hidden size in FFN. f: minimum token frequency in vocabulary. M: context length. L: number of encoder layers. h: number of heads in MultiHead. λ_1, λ_2: weight coefficients of $loss_2, loss_3$, repectively.

5 Result Analysis

Model	EC	DailyDialogue	MELD	EmoryNLP	IEMOCAP
cLSTM	69.13	49.90	49.72	26.01	34.84
CNN (Kim, 2014)	70.56	49.34	55.86	32.59	52.18
CNN+cLSTM (Poria et al., 2017)	72.62	50.24	56.87	32.89	55.87
Bert_BASE (Devlin et al., 2019)	69.46	53.12	56.21	33.15	61.19
DialogueRNN (Majumder et al., 2019)	74.05	50.65	56.27	31.70	61.21
KET (Zhong et al., 2019b)	73.48	53.37	58.18	34.39	59.56
KAITML (ours)	**75.39**	**54.71**	**58.97**	**35.59**	**61.43**

Table 4: Performace comparisons on the five test sets (%). Bold font denotes the best performance.

5.1 Comparison with Baselines

Table 4 shows the performance of different models on 5 benchmark datasets. We can see that our model outperforms all the baselines, on all the datasets, which shows the effectiveness of our proposed model

for ERTC. Through paired t-test, there were significant differences between our proposed model and all baselines ($p \leq 0.05$). Note that all the results of baselines are directly cited from (Zhong et al., 2019b).

The state-of-the-art KET model performs best overall among all baselines. And our KAITML model surpasses the KET model by around 1.5% performance on all the dataset tested. To explain this gap in performance, it's significant to understand the nature of these models. KAITML and KET both incorporate external commonsense knowledge and model contextual information based on transformer for ERTC. This is a key limitation in other baseline models, as external commonsense knowledge can enrich the background and semantic information of utterances and the self-attention module in transformer allows model to exploit contextual information more efficiently than CNNs and gated RNNs in other baseline models. As for the difference of performance between KAITML and KET, we believe that this is due to the difference of graph attention mechanism and multi-task learning. That KET doesn't consider various relations in external knowledge base may cause the loss of semantic information. By contrast, KAITML tries to overcome this issue by using a dual-level graph attention mechanism, which can exploit the various relations in external knowledge base and thus support better understanding of utterances. In addition, the multi-task learning in KAITML can alleviate the confusion between a few non-neutral utterances and much more neutral ones and thus further improve the emotion recognition performance.

5.2 Ablation Study

context	knowledge	multi-task	EC	DailyDialogue	MELD	EmoryNLP	IEMOCAP
✓	✓	✓	**74.97**	**56.76**	**54.22**	**38.10**	**50.83**
✗	✓	✓	68.92	54.84	52.68	37.85	49.25
✓	✗	✓	73.90	55.87	53.09	34.75	49.16
✓	✓	✗	73.46	55.57	53.83	36.84	49.19

Table 5: Ablation results on five validation sets (%). Context, commonsense knowledge and multi-task learning are all beneficial to the emotion recognition performance.

To comprehensively study the impact of context, knowledge and multi-task learning, we remove them one at a time and investigate their contribution on all datasets. As expected, following Table 5, context, knowledge and multi-task learning are all essential to the strong performance of our model on all datasets and their combination achieves the best performance. Note that removing knowledge has a greater impact on small datasets (i.e., EmoryNLP and IEMOCAP) than big datasets (i.e., EC, DailyDialogue and MELD), which is expected because external commonsense knowledge can help model understand utterances, especially when there is insufficient data. Moreover, compared to other datasets, the performance of the EC drops a lot, around 6%, when removing context. The reason may be that there are more short utterances on EC, like "ok", "yes", whose emotion depends on the context it appears in. With multi-task learning, we observed that the confusion between non-neutral categories and neutral category is alleviated in the confusion matrix and the performance improves by about 1.2% on all datasets on average.

5.3 Error Analysis

By analyzing our predicted emotion labels, we found that the model error is mainly caused by the following aspects. Firstly, misclassifications are often among similar emotion classes in the confusion matrix, like 'happy' and 'excited', 'angry' and 'frustrated'. Secondly, the performance of emotion classes with small amount data available is poor, like 'fear' and 'disgust' in DailyDialogue dataset. Thirdly, some of datasets, such as MELD, IEMOCAP, EmoryNLP, that we use in our experiment are multimodal. And we found that acoustic, and visual modality provide key information to recognize emotions in a few utterances (e.g., 'okay', 'yes', etc.) while our proposed KAITML model considers only textual modality.

6 Conclusion and Future Work

In this paper, we propose a novel Knowledge Aware Incremental Transformer with Multi-task Learning (KAITML) for emotion recognition in textual conversations, which can effectively incorporate contextual information and commonsense knowledge, and alleviate the confusion between a few non-neutral utterances and much more neutral ones. Moreover, extensive experimental results show that our KAITML model outperforms state-of-the-art models across five benchmark dataset. Future work will focus on the following directions: 1) how to differentiate similar emotions, 2) how to recognize emotion using limited data, 3) how to incorporate multimodal information for emotion recognition in conversations.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Key Programs of the Chinese Academy of Sciences (ZDBS-SSW-JSC006) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No.XDB32070000).

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding Emotions in Text Using Deep Learning and Big Data. *Comput. Hum. Behav.*, 93:309–317.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019. A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2687–2693.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 551–561.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 221–231.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132.

Junqing He, Bing Wang, Mingming Fu, Tianqi Yang, and Xuemin Zhao. 2019. Hierarchical Attention and Knowledge Matching Networks With Information Enhancement for End-to-End Task-Oriented Dialog Systems. *IEEE Access*, 7:18871–18883.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Chenyang Huang, Osmar R. Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic Dialogue Generation with Expressed Emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 49–54.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 12–21.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1468–1478.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 821–832.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019b. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7:100943–100953.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion Detection in Text: a Review. *CoRR*, abs/1806.00674.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2018. How Time Matters: Learning Time-Decay Attention for Contextual Spoken Language Understanding in Dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2133–2142.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Sida I. Wang and Christopher D. Manning. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 90–94.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, pages 44–52.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 533–542.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019a. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7492–7500.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019b. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 165–176.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629.