AutoML-Emo: Automatic Knowledge Selection using Congruent Effect for Emotion Identification in Conversations

Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria

Abstract—Emotion recognition in conversations (ERC) has wide applications in medical care, human-computer interaction, and other fields. Unlike the general task of emotion analysis, humans usually rely on context and commonsense knowledge to convey emotions in conversations. Only when the model can connect and fully utilize a large-scale commonsense knowledge base, it can better understand latent contents in conversations. Unfortunately, there is no available knowledge selection mechanism to address such knowledge needs and to make sure the system is not flooded with irrelevant commonsense knowledge. Therefore, we propose an AutoML strategy based on emotion congruent effect to select suitable knowledge and models, called AutoML-Emo. Global exploration and local exploitation-based selection mechanism (G&LESM) are used for automatic knowledge selection. The transformer-based architecture search (TAS) is applied to model selection, the selected transformer-based model is employed to incorporate knowledge and capture context information in conversations. The experimental results show that AutoML-Emo can effectively enhance external knowledge in different sizes and domain datasets. Moreover, the selected transformer-based model derived from TAS is superior to the most advanced models.

Index Terms—Autonomous machine learning, Genetic algorithm, Knowledge selection, Emotion recognition.

1 Introduction

Emotional recognition in conversations (ERC) received widespread attention from researchers [1], [2], [3] recently. When people talk to each other, humans make commonsense inferences to determine their understanding of the narrative being presented [4]. In addition, humans often rely on context and commonsense knowledge to convey emotions [5], which makes the machine hard to recognize and understand the emotions of utterances unless it can connect and fully utilize the huge knowledge base [6]. Therefore, knowledge selection has meaningful implications for ERC. In recent work, Poria et al. [7] used a recurrent neural network (RNN) to model contextual utterances in order of time, in which each utterance is represented by a feature vector. Majumder et al. [8] combined the attention mechanism to gather the information of each target utterance. Hazarika et al. [9] proposed a memory network to model context. However, these methods only focus on context and do not utilize external commonsense knowledge to recognize and understand emotions. Incorporating commonsense knowledge from the external knowledge base is the basis for understanding the content of conversations and making empathic responses [10], [11].

- *D. Jiang is with the Department of Computer Science, Shantou University, Shantou, China, and the Guangdong Provincial Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, China (E-mail: dzjiang@stu.edu.cn); R. Wei, J. Wen and G. Tu are with the Department of Computer Science, Shantou University, Shantou, China. (E-mail: 20rgwei@stu.edu.cn; 20jtwen@stu.edu.cn; 19gtu@stu.edu.cn).*
- *E. Cambria. is with School of Computer Science and Engineering, Nanyang Technological University, Singapore. (E-mail: cambria@ntu.edu.sg).*
- *E. Cambria is the corresponding author.*

Therefore, Zhong et al. [5] applied a knowledge-enriched transformer (KET) to enrich the semantics of utterances by referring to knowledge entities from external knowledge bases. Zhou et al. [12] employed a graph convolutional network to learn the representation of relevant knowledge. In addition, Ghosal et al. [6] proposed a new framework, called COSMIC, for incorporating different commonsense elements such as events, mental state, and so on. Obviously, these models do not have an available knowledge selection mechanism. External knowledge not only enhances the text semantics but also brings a lot of noise to data, especially when the size of data is large. If we search for the best combination of knowledge manually, it will greatly increase the cost of the experiment. In addition, the huge knowledge shows the complexity of the knowledge selection task. Fig. 1 shows an example in conversations, which illustrates the importance of knowledge selection in recognizing and understanding the emotions of utterances.

In other words, incorporating commonsense knowledge is becoming increasingly popular in ERC [13], [14], but it also brings in a lot of noise to data. AutoML is a strategy that automatically searches for a suitable combination of parameters, algorithms, and so on [15], [16]. Inspired by such methods, an automatic knowledge selection mechanism is composed of AutoML and commonsense knowledge selection has become a topic worth exploring. This ERC task and its extensive experiment can be better handled by the method based on AutoML strategy [17]. Consequently, we propose an AutoML strategy called AutoML-Emo, which can search automatically for the suitable commonsense knowledge combination in the massive external knowledge base and the appropriate transformer-based model derived from transformer-based architecture search (TAS).

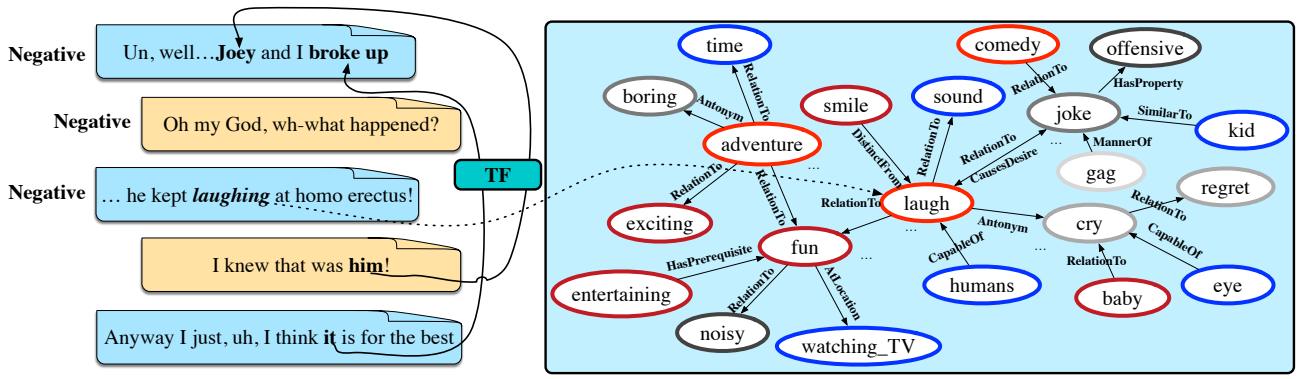


Fig. 1. A example of a conversation. (Legend: TF stands for the transformer-based model. Additionally, we can find that "him" and "it" in the fourth and fifth utterance are related to "Joey" and "break up" in the first utterance due to referring to the contextual information. The "laughing" in the third utterance is enriched by referring to the external knowledge base. Especially, the different color stands for different sentiment such as red blue and gray represents positive neutral and negative, respectively. The depth of the color denotes emotional intensity.)

The transformer-based model [18] has been shown to be a powerful representation learning model in many NLP tasks, such as machine translation[18] and ERC [5]. Experimentally, it was found that changes in parameters have a significant impact on performance due to the lack of an automatic learning process for the transformer-based model parameters. Specifically, the AutoML-Emo framework is essentially a process of knowledge selection, which can be divided into the following two parts: global exploration and local exploitation-based selection mechanism (G&LESM). The global exploration selection mechanism (GESM) is a knowledge selection mechanism based on emotion congruent effect [19], which is used to quickly and globally select appropriate knowledge from unknown massive knowledge, thus reducing the scale of knowledge. In fact, the recalled information may be affected by the change of emotions, which is called the emotion congruent effect. GESM is a selection algorithm based on grid search, which means that it sacrifices a certain precision to improve search efficiency. And its selection strategy is in line with the emotion congruent effect, which offers a realistic basis and certain interpretability. As a result, GESM will not only consider the relationship between commonsense knowledge and corresponding words but also focus on selecting commonsense knowledge consistent with the sentiments of words. Additionally, sentimental intensity is used to depict the degree of sentimental consistency.

The local exploration selection mechanism (LESM) is a selection mechanism based on the genetic algorithm [20], which is applied to accurately and locally select the knowledge gained by GESM, thus improving the quality of introduced knowledge. Especially, when positive words appear in a negative context, GESM has obvious limitations, while LESM only selects knowledge based on fitness function (the validation loss), which can make up for the deficiency of GESM. Inspired by the emotion congruent effect, the selection of knowledge has new theoretical guidance, and the selected knowledge will be given more interpretability. The TAS allows the selected model to be better adapted to the selected knowledge entities, through referring to these selected knowledge entities, the selected transformer model derived from TAS is superior to most advanced models.

From the view of experimental results, the selected knowledge can help the selected model better adapt to different sizes and domain datasets. The main contributions of this paper are summarized as follows:

1) We propose a GESM based on emotion congruent effect, which can quickly and globally select the unknown huge knowledge, so as to reduce the size of commonsense knowledge and pave the way for further selection. And it will tend to search the knowledge entitled consistent with the sentiments of words, in which the sentimental intensity is used for describing the degree of emotional congruence.
2) We propose a LESM based on a genetic algorithm, which can accurately and locally select the knowledge obtained by GESM, so as to effectively improve the quality of external knowledge. In particular, it can make up for the deficiency of GESM, that is, the limitation of positive words in a negative context when selecting knowledge.
3) We conduct a lot of experiments and find that using AutoML-Emo to select commonsense knowledge is helpful to ERC. In addition, on the different sizes and domain datasets, the knowledge we selected makes the performance of the transformer-based model derived from TAS better than the most advanced model.

The rest of this paper is organized as follows: Section 2 illustrates related work; Section 3 introduces AutoML-Emo; Section 4 and 5 list the extensive experiments and analysis conducted to show the effectiveness of our proposed method; finally, Section 6 offers to conclusion and future directions.

2 Related Work

Emotion analysis around conversations is an important topic in recent years, which has attracted much attention in natural language processing. The availability of many conversation datasets [21], [22], [23] partly explain this phenomenon, and the growing interest in conversation systems can also explain this phenomenon [24], [25].

Emotion recognition in conversations: Early research in ERC mainly employed vocabulary-based methods [26], [27], [28]. In recent years, researchers in ERC began to adopt deep learning technology [29], [30], [31], [32], [33]. For example, Poria et al. [34] proposed a contextual long-short term memory to capture the context information. Hazarika et al. [9] proposed a gated recurrent unit (GRU) based on a conversational memory network (CMN), which builds different contextual models for the speakers and the listeners. Additionally, a DialogueRNN model was proposed in [8], which used three GRUs to model the emotional states, the context, and the speaker states. Ghosal et al. [35] proposed a method based on a graph neural network, which uses the dependency relationship between the different speakers to model the context.

External knowledge in conversations: Commonsense knowledge is the foundation for understanding conversations and generating appropriate responses, which can lead to explainable dialogue understanding [5], [6]. External commonsense knowledge is a collection of commonsense knowledge [10]. In textual conversations, there is a lot of knowledge that is obvious to humans but difficult to be recognized by models. For example, speaker A: "I like dogs best, and you?", speaker B: "I like Collie best.". It is difficult for a machine to conclude that "Collie" is a "dog" from the contextual utterances unless a connection is established between "Collie" and "dog". To address this problem, we have to incorporate knowledge from external large-scale knowledge bases, such as ConceptNet [36] and SenticNet [37]. Young et al. [38] proposed the first end-to-end dialogue system augmented with commonsense knowledge. Zhong et al. [5] proposed a knowledge-enriched transformer, which uses context-aware graph attention to embed commonsense knowledge. Zhang et al. [39] introduced a dual-level graph attention to fuse external knowledge for enhancing the semantics of target utterances. Ghosal et al. [6] proposed a new framework, called COSMIC, which incorporates commonsense elements and uses it as the basis to learn the dependency between interlocutors. In addition, external commonsense knowledge is a collection of commonsense knowledge [10].

AutoML in emotion analysis: Autonomous machine learning (AutoML) [15] focuses on developing an effective method for automatically designing machine learning workflow, which does not require a lot of human intervention [40]. Recently, there are some researchers have begun to employ AutoML for emotion analysis. For example, Lopes et al. [41] proposed a fusion classification method based on AutoML, which combines text and image sentiment analysis and finds the best model through a random search strategy. AL-Sharuee et al. [42] introduced an automatic and unsupervised sentiment analysis method to analyze comment sentiment. Chen et al. [43] introduced a new lifelong learning emotion classification method, which uses a Bayesian optimization framework based on random gradient descent. Unfortunately, there is no work related to AutoML in ERC, let alone the combination of AutoML and external knowledge in conversations.

3 METHODOLOGY

In this section, we propose an AutoML strategy based on emotion congruent effect, which is used to select suitable knowledge and transformer-based model. It consists of three parts shown in Fig. 2: GESM, LESM, and TAS. After GESM based on congruent effect, the scale of external knowledge related to "laughing" is reduced because the negative and neutral knowledge is deleted when facing a negative context. However, LESM makes up for the defects of GESM by further reducing the positive knowledge according to the fitness function. In the process of TAS, we use the internal parameter spaces in the transformer-based model as the search space.

3.1 Task Definition

Let the hyperparameter spaces related to the knowledge selection and the transformer-based architecture search be $\Lambda = \{\Lambda^1, ..., \Lambda^n\}$ and $\bar{\Lambda} = \{\bar{\Lambda}^1, ..., \bar{\Lambda}^n\}$, respectively. Let $\{u_j, c_j^{(i)}, y_j\} \in \{U, C, Y\}$ denotes the tuple of utterance, knowledge, label, where $j = 1, ..., N$ represents number of utterances, $c_j^{(i)}$ stands for the ith commonsense knowledge in the jth utterance, and y_j is the emotional label of jth utterance (see Fig. 2). Additionally, U, C, and Y represent all utterances and the set of their corresponding knowledge and emotion labels. Thus, the knowledge selection problem can be written as:

$$\theta^* = \underset{\theta \in \Lambda}{argmin}\ {}_{(u_{train/valid}, c_{train/valid}) \sim (U,C)} \mathcal{L}^*(\Lambda, (u_{train}, c_{train}), (u_{valid}, c_{valid}), \lambda^*) \quad (1)$$

where $\lambda \in \bar{\Lambda}$ denotes the hyperparameter spaces related to model $\mathcal{M}$, and λ^* represents a given combination of hyperparameters. In knowledge selection, the model $\mathcal{M}^*$ used to evaluate performance is not in the scope of TAS. Thus, by default, model $\mathcal{M}^*$ is initialized with the given λ^*. The $\mathcal{L}^*(\Lambda, (u_{train}, c_{train}), (u_{valid}, c_{valid}), \lambda^*)$ is the loss function when model $\mathcal{M}^*$ is trained on (u_{train}, c_{train}) and evaluated on (u_{valid}, c_{valid}). In addition, the transformer-based architecture search problem can be written as:

$$\lambda^{**} = \underset{\lambda \in \bar{\Lambda}}{argmin}\ {}_{\mathcal{M}_\mathcal{L}} \mathcal{L}(\theta^*, (u_{train}, c_{train}), (u_{valid}, c_{valid}), \lambda) \quad (2)$$

where $\theta^* \in \Lambda$ is the combination of hyperparameters from G&LESM, and λ^{**} is the suitable hyperparameters combination of model $\mathcal{M}$ from TAS. The $\mathcal{L}(\theta^*, (u_{train}, c_{train}), (u_{valid}, c_{valid}), \lambda)$ is the loss function of model $\mathcal{M}$. Additionally, in model $\mathcal{M}$, we limit the size of context windows to $\mathcal{M}$, thus reducing the calculation cost in the evaluation process. Discarding the early contextual utterances may lead to a negative impact, but it is negligible because they only contribute the least information [44].

3.2 Global exploration-based selection mechanism (GESM)

Incorporating external knowledge can enhance the semantics of utterances but while playing a positive role, they also bring a lot of noise to data, so an effective knowledge selection way is quite urgent. However, in the face of such vast and complex knowledge, an accurate knowledge selection strategy will bring huge experiment costs, and the results searched are easy to fall into local optima.

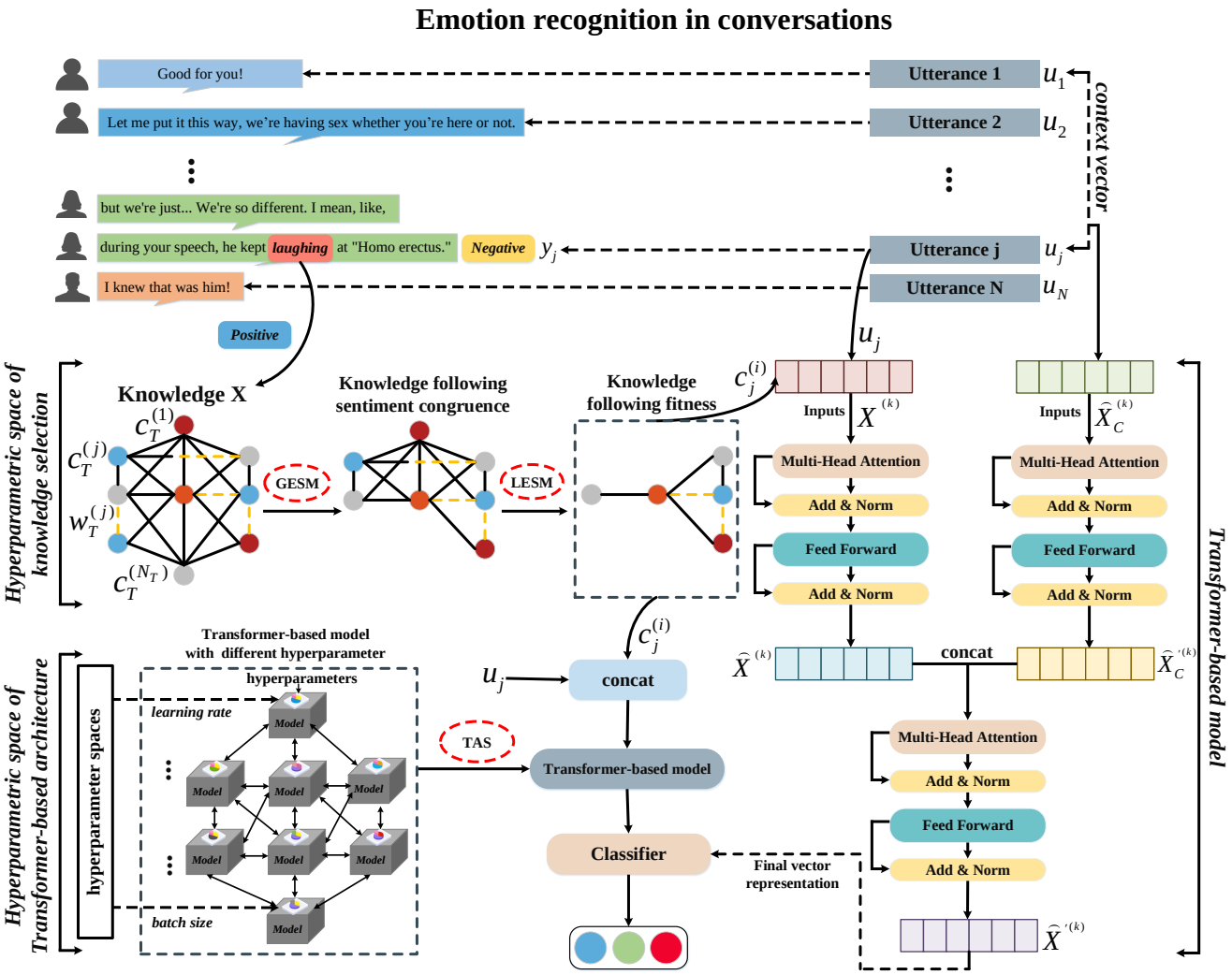


Fig. 2. The overall architecture of AutoML-Emo framework. (Legend: Where $c_j^{(i)}$ stands for the ith commonsense knowledge in the jth utterance, and y_j is the emotional label of jth utterance. In the hyperparametric of Transformer-based architecture, $X^{(k)}$ is the representation of the kth utterance $u^{(k)}$ concatenated with the corresponding knowledge. $\widehat{X}^{(k)}$ is the vector representation of $X^{(k)}$ processed by MSAT, and $\widehat{X}_C^{\prime(k)}$ represents the context vector representation of $\widehat{X}^{(k)}$. $\widehat{X}^{\prime(k)}$ represents the final vector representation of $u^{(k)}$ and corresponding commonsense knowledge, which is as input to the subsequent fully connected network for emotion classification.)

Additionally, the process of selecting knowledge like violent search has no realistic basis, which lacks interpretability is also a big problem. In order to solve these problems, a fast, effective, low-cost automatic selection mechanism GESM is introduced. Moreover, in the process of selecting knowledge, the GESM not only reduces the scale of knowledge, paves the way for local and more accurate search, but also ensures that the selected knowledge will not lose sentimental consistency with the target words.

$$C_T = \sum_{j=1}^{|N_T|} {W_T}^{(j)} \cdot c_T^{(j)} \tag{3}$$

where ${W_T}^{(j)}$ represents the weight of the jth knowledge $c_T^{(j)}$ associated with the target word T. N_T is the total number of knowledge. C_T stands for the vector representation of related knowledge, which is a common method to obtain knowledge representation.

The goal of GESM is not to select an optimal way of knowledge representation, but to quickly select W_T suitable for different sizes and domain data from a global perspective. The specific calculation process is as follows:

$$\widehat{W}_T^{(j)} = \begin{cases} \varepsilon \to 0, & S_j \geq \Omega^* \ and \ Senti(T) \neq Senti(c_T^{(j)}) \\ W_T^{(j)} \cdot \mathbb{Z}(c_T^{(j)}), & Senti(T) = Senti(c_T^{(j)}) \end{cases} \tag{4}$$

$$\widehat{W}_T^{(j)} = \eta \cdot W_T^{(j)} + (1-\eta) \cdot \widehat{W}_T^{(j)} \tag{5}$$

$$S_j = R^{1/W_T^{(j)}}, R \sim U(0,1) \tag{6}$$

$$\mathbb{Z}(c_T^{(j)}) = (\left\| V(c_T^{(j)}) - 0.5 + \mathrm{A}(c_T^{(j)}) / 2 \right\|_2 - \alpha) / \beta \tag{7}$$

where $\widehat{W}_T^{(j)}$ represents the weight between target word T and corresponding knowledge, which is the result of ${W_T}^{(j)}$ treated by GESM. $Senti(\cdot)$ is a sentiment recognition method based on CoreNLP [45]. ε is a small number closing to zero. S_j is a weighted random number. Because the initial

weights of knowledge are different, the random number generated by uniform distribution can not reflect the difference. $\mathbb{Z}(\cdot)$ a method to measure sentimental intensity. Additionally, $V(\cdot)$ and $A(\cdot)$ denote the value of knowledge in the dimension of valence and arousal, in the valence-arousal-dominance (VAD) emotion model [46]. Noticeably, we use sentimental intensity to further depict the degree of emotional consistency between target words and corresponding knowledge. And $\Omega^* \in \Lambda$ is the combination of hyperparameters related to GESM. Its optimization process can be written as:

$$\Omega^* = \underset{\Omega \in \Lambda}{argmin}\, \mathcal{M}_{\mathcal{L}} \mathcal{L}(\Lambda, (u_{train}, \widehat{c}_{train}), (u_{valid}, \widehat{c}_{valid}), \lambda^*) \quad (8)$$

where $\widehat{c}_{train/valid}$ represents the set of knowledge representations corresponding to all target words, which can be obtained according to weight vector $\widehat{W}_T \in {}^{N_T \times N_C}$. N_T and N_C represent the total number of different target words and corresponding knowledge, respectively. Ω indicates the possible value range of the hyperparameters in a grid search. The goal of GESM is to find the best result in a given range, which not only satisfies the emotional consistency between the target word and the corresponding knowledge, but also boosts the performance of the evaluation model. The pseudocode for GESM is given in Algorithm 1.

Algorithm 1 Global exploration-based selection mechanism

initialize model $\mathcal{M}^*_{\mathcal{L}^*}; \mathcal{H} \leftarrow \varnothing$;
initialize hyperparameter $\lambda^*; \Omega \leftarrow \{\Omega_{n_s} = \Omega_1 + (n_s - 1) \cdot d\}$;
initialize variable $\ell_{\min} \leftarrow \infty$;
for Ω_k in hyperparameter space Ω **do**
  compute new knowledge representation $\widehat{c}_{\text{train/valid}}$ ▷ equation (3)
  evaluate model $\mathcal{M}^*_{\mathcal{L}^*}$ on $(u_{\text{valid}}, \widehat{c}_{\text{valid}})$
  record F1 score ℓ_k of $\mathcal{M}^*_{\mathcal{L}^*}$
  if $\ell_k < \ell_{\min}$ **then**
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\Omega_k, \ell_k, \lambda^*)\}$
    repeat
      update $\mathcal{M}^*_{\mathcal{L}^*}$ given $\mathcal{H}$
    until stopping criterion met
  end if
end for
return Ω^* from $\mathcal{H}$ with maximal ℓ

3.3 Local exploitation-based selection mechanism (LESM)

After GESM, the scale of external knowledge has been greatly reduced, but the quality of knowledge has been upgraded. However, the random selection of knowledge only from the perspective of sentiment is not enough for different sizes and domain datasets. What's more, GESM is just a search strategy, which sacrifices a certain precision for search efficiency. Therefore, we propose a LESM to make a further and more accurate selection on GESM results. Unlike GESM, LESM applies fitness function (the validation loss) as the selection basis to search the existing results accurately. Although GESM can also play an active role in the experimental results, GESM also has certain limitations when positive words appear in a negative context. At this time, LESM makes up for the shortcomings of GESM.

Algorithm 2 Local exploitation-based selection mechanism

Require:
$\mathcal{P}_c$ -the crossover Probability ;
$\mathcal{P}_m$ - the mutation Probability ;
$\mathcal{D}$ - the population size ;
$\mathcal{G}$ - the number of generations
Ensure:
v^* - the optimal chromosomes ;
initialize variable $\varphi \leftarrow \varnothing$
create initial population $\mathcal{V}$ ▷ Algorithm 3
repeat
  compute the fitness of each individual $\mathcal{F}_k$ in the population
  initialize empty population $\varphi \leftarrow \varnothing$
  repeat
    $\{\alpha, \beta\} \leftarrow$ select operation to $\mathcal{V}$ according to $\mathcal{F}$ section 3.3
    if $\text{random}(0,1) < \mathcal{P}_c$ **then**
      crossover operation to $\{\alpha, \beta\}$
    end if
    if $\text{random}(0,1) < \mathcal{P}_m$ **then**
      mutation operation to $\{\alpha, \beta\}$
    end if
    $\varphi \leftarrow \varphi \cup \{\alpha, \beta\}$
  until $\mathcal{D}$ offsprings were created
  $\mathcal{V} \leftarrow \varphi$
until reproductive generations over $\mathcal{G}$
return v^* from $\mathcal{V}$ with minimal fitness

The LESM is a search strategy based on a genetic algorithm, and the pseudocode for LESM is given in Algorithm 2. The genetic algorithm (GA) is a random search algorithm based on genetic mechanisms and natural selection. It mainly consists of the following components: coding mechanism, fitness function, genetic operator (such as crossover and mutation), and control parameters. Thus, the LESM can be introduced as follow:

Coding mechanism and initialization population: When GA is used to solve problems, the possible solution needs to be coded as a chromosome, that is, an individual and several individuals form an initial solution group. Therefore, we regard all the non-zero values in $\widehat{W} \in {}^{N_T \times N_c}$ as 1, and then expand them into a binary vector $v \in {}^{N_V}$ where N_V is the length of the chromosome. Finally, we generate the initial individuals according to the weighted random method and form the initial solution group $\mathcal{V} \in {}^{N_G \times N_V}$ The pseudocode is given in Algorithm 3.

Algorithm 3 Creation of Initial Population

initialize variable $\mathcal{V} \leftarrow \varnothing$
for each $\mathcal{G} \leftarrow \{1, \cdots, N_G\}$ **do**
  $\mathcal{S} \leftarrow \{1/S_1, \ldots, 1/S_{N_V}\}$ ▷ equation (6)
  $\tau \leftarrow 0;\ v \leftarrow \{0\} * N_V$
  for ξ in $\mathcal{S}$ **do**
    $\psi \leftarrow (\xi - \xi_{\min}) / (\xi_{\max} - \xi_{\min})$ max-min normalization
    if $\psi > 0$ **then** $v_\tau \leftarrow 1$
    end if
  end for
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{v\};\ v \leftarrow \{0\} * N_V$
end for
return $\mathcal{V}$

Fitness function and termination condition: To make a genetic algorithm measure the superiority of individuals in the population, a fitness function must be defined. Here,

similar to the GESM approach, by default, the loss function of transformer-based model $\mathcal{M}^*$ is used as the fitness function to evaluate the fitness of each chromosome in the population. The specific calculation process of the loss is as follows.

$$\mathcal{L}(y, \widehat{y}) = \sum_{k} W_F[-y_k + log(\sum_{j=1}^{N_e} e^{\widehat{y}_k[j]})] + \lambda \left\|\theta\right\| \quad (9)$$

where y and $\widehat{y}$ respectively represent the true and predicted emotional labels of each sentence $u^{(k)}$, on validation sets. N_e indicates the total number of emotion categories. W_F is learnable parameters of the transformation. The λ is the L2 regularization term and θ is the set of W_F and other learnable parameters of the transformation in transformer-based model $\mathcal{M}^*_{\mathcal{L}^*}$. In each iteration of GA, the individuals in the current population are evaluated and ranked according to the fitness function. Among them, individuals with lower fitness will be more likely to survive into the next generation or mating pool. When the number of generations reaches $\mathcal{G}$. The individual with the least-loss is selected and then the algorithm is terminated.

Genetic operator and generation of new population: After fitness evaluation, the algorithm uses genetic operators to create a new population. In this part, we mainly introduce the crossover and mutation operators of LESM. **Crossover operation:** the crossover operator in a genetic algorithm combines two individuals to form the offspring of the next generation. The two parent chromosomes needed for crossover operation are the two individuals with the least-loss selected according to the tournament algorithm. Through crossover, the searchability of the genetic algorithm is improved. In LESM, the crossover of two chromosomes is realized by the XOR operation, as shown below.

$$CrossOverKids(\mathcal{I}) = P_1 \oplus P_2 \quad (10)$$

where $\mathcal{I}$ is an index that runs from 1 to the number of kids, and $\oplus$ is an XOR operator for binary operands. The P_1 and P_2 are the first and second parent, respectively, which are needed by the crossover operator. **Mutation operation:** mutation is the genetic interference of individuals in a population. Mutation ensures genetic diversity and search for wider solution space. The LESM adopts uniform mutation, that is, a genetic algorithm generates a random number set of genome lengths from a uniform distribution. The value of each random number is related to the position of each gene on the chromosome. The chromosome is scanned from left to right, and the value of $\mu^{(k)}$ is compared with the mutation probability $\mathcal{P}_m$ for each associated bit k. if the $\mu^{(k)}$ at position k is less than $\mathcal{P}_m$, the gene (bit) at position k is flipped. Otherwise, the gene at position k would not be affected. **New population:** In this way, the genetic algorithm has been evolving until the new population is filled. The new population is filled by increasing the number of elite kids, cross kids and mutant kids. Among them, elite kids are the two least-loss chromosomes in the previous generation. They can directly survive into the next generation or mating pool without crossover and mutation.

3.4 Transformer-based architecture search (TAS)

In this section, our approach is similar to neural architecture search (NAS) [47], but instead of the whole process of building a machine learning workflow, we use the internal parameter spaces in the transformer-based model as the search space. Among them, the transformer-based model shown in Fig. 2 is employed to incorporate knowledge and capture context information in conversations. The specific calculation process is as follows:

$$\widehat{X}^{(k)} = PN(MSAT(X^{(k)} = [u^{(k)};\ \widehat{c}^{(k)}],\ X^{(k)}, X^{(k)})) \quad (11)$$

$$\widehat{X}^{\prime(k)} = PN(MSAT(\widehat{X}^{(k)},\ \widehat{X}_C^{\prime(k)},\widehat{X}_C^{\prime(k)})) \quad (12)$$

$$MSAT(Q, K, V) = soft\max(\frac{Q\,K^T}{\sqrt{d_s}})\,V \quad (13)$$

$$PN(x) = max(0,\ xW_P^{(1)} + b_P^{(1)})W_P^{(2)} + b_P^{(2)} \quad (14)$$

where $X^{(k)}$ is the representation of the kth utterance $u^{(k)}$ concatenated with the corresponding knowledge. $PN(\cdot)$ and $MSAT(\cdot)$ represent Position-wise feed-forward networks (PN) and Multi-head self-attention mechanism (MSAT), respectively. $\widehat{X}^{(k)}$ is the vector representation of $X^{(k)}$ processed by MSAT, and $\widehat{X}_C^{\prime(k)}$ represents the context vector representation of $\widehat{X}^{(k)}$. $\widehat{X}_C^{\prime(k)}$ is the vector representation of $\widehat{X}_C^{(k)}$ processed by MSAT. $\widehat{X}^{\prime(k)}$ represents the final vector representation of $u^{(k)}$ and corresponding commonsense knowledge, which is as input to the subsequent fully connected network for emotion classification. $W_P^{(1)}$, $W_P^{(2)}$ are learnable parameters of the transformation. $b_P^{(1)}$ and $b_P^{(2)}$ are bias values of PN. To find the optimal combination of hyperparameters and algorithms, based on tune [48], we automatically and randomly search a group of machine learning algorithms and their internal parameters [49]. Therefore, in this work, we randomly search the internal parameter spaces containing the following: loss function, learning rate, batch size, optimizer, the number of heads in multi-head attention mechanism, the number of hidden layers in position-wise feed-forward networks, the size of word embedding (d), the size of context windows (M). Finally, the best combination of hyperparameters and algorithms is selected to make the best performance of the model in the validation set. Especially, we initialize the words and knowledge in conversations by Glove embedding [50].

4 Experiment

In this section, we conduct experiments to verify the effectiveness of automatic knowledge selection mechanism AutoML-Emo, on MELD [21], DailyDialog [22] and EmoryNLP [23] datasets.

4.1 Datasets

We test AutoML-Emo on three different conversational datasets. **DailyDialog:** humans daily written communication. **MELD and EmoryNLP:** TV program scripts collected from "Friends". However, the size and annotation of EmoryNLP are different from MELD. The emotion labels of EmoryNLP include neutral, sad, mad, scaled, powerful, peaceful, and joyful. Additionally, in terms of evaluation indicators, for DailyDialog, we use micro F1 according to [51] because their labels are extremely unbalanced (the percentage of the main categories in the test set is more

TABLE 1
The splits and evaluation metrics used in different datasets.

Dataset	# Dialogue			# Utterances			# Class	# Metric
	Train	Val	Test	Train	Val	Test		
MELD	1,039	114	280	9,989	1,109	2,610	7	Weighted Avg.F1
DailyDialog	11,118	1,000	1,000	87,832	7,912	7,863	7	Micro F1
EmoryNLP	659	89	79	7,551	954	984	7	Weighted Avg.F1

TABLE 2
The detailed hyperparameters setting of GESM.

Dataset	$\#\ \Omega^*$	$\#\ \alpha$	$\#\ \beta$	$\#\ \eta$
MELD	0.1	0.06467	0.607468	0.5
DailyDialog	0.6	0.06467	0.607468	0.5
EmoryNLP	0.3	0.06467	0.607468	0.5

TABLE 3
The detailed hyperparameters setting of LESM. (Legend: elite kids are the two least-loss chromosomes in the previous generation. They can directly survive into the next generation or mating pool without crossover and mutation.)

# Hyperparameter	# Value
Population size	100
Fitness Function	loss of transformer-based classifier
Number of generations	300
Crossover	arithmetic crossover
Crossover Probability	0.8
Mutation	uniform mutation
Mutation Probability	0.1
Selection scheme	tournament of size 2
Number of elite kids	2

than 80%). For the other relatively balanced datasets, we use weighted avg.f1 following [8]. The more information about datasets is shown in Table 1 below.

4.2 Commonsense Knowledge

The external knowledge bases such as SenticNet and ConceptNet are applied in this paper. Emotion dictionary NRC_VAD [52] is the source of sentimental intensity in our model. **ConceptNet:** a semantic network in which each word and phrase are connected to each other by labeled (representing the type of edge) and weighted (representing the credibility of edge). **SenticNet:** a knowledge base, which provides a set of 200,000 natural language concepts related to semantics, emotion and polarity. In particular, emotion refers to the emotional value of the four emotional dimensions (pleasant, attention, sensitivity, and aptitude) in hourglass model [53] and the emotional polarity value between -1 and +1 (where -1 is extremely negative and + 1 is extremely positive). **NRC_VAD:** an emotional dictionary, which contains a list of english words and their scores, that is, the scores of arousal, valence and dominance in the [0,1] interval.

4.3 Baseline Methods

In this section, we introduce some baselines in ERC. **CLSTM [31]:** a utterance-level bidirectional LSTM is used to encode each sentence. **CNN [54]:** a single-layer CNN with strong empirical performance. which is trained in the context-free utterance level. **BERT_BASE [55]:** the basic version of the latest model of emotion classification. It treats each utterance and its context as a separate document and limits the document length to the last 100 tags to allow a larger batch size. Because of the memory limitation of GPU, we don't use the large version of Bert for the experiment. **DialogueRNN [8]:** it models the emotional state, context, and speaker state in conversations via three GRU networks. **Att-NDE [56]:** it presents a new continuous-time attention method for sequential learning which is tightly integrated with NDE to construct an attentive continuous-time state machine. **Co-GAT [57]:** it designs a co-interactive graph attention network to model simultaneously incorporate contextual information and mutual interaction information. **HiTransformer [58]:** it proposes a hierarchical transformer framework with a lower-level transformer to model the word-level input and an upper-level transformer to capture the context of utterance-level embeddings. **KET [5]:** it uses hierarchical self-attention and context-aware graph attention to incorporate external commonsense knowledge dynamically. **AutoML-Emo (ours):** a transformer-based model, which can be used to capture the context and incorporate commonsense knowledge. In addition, its external knowledge can better adapt to different sizes and domain datasets, after GESM and LESM.

4.4 Hyperparameter Settings

The setting of related hyperparameters and algorithms involved in the process of knowledge selection and transformer-based architecture search is introduced as shown in Table 2-4.

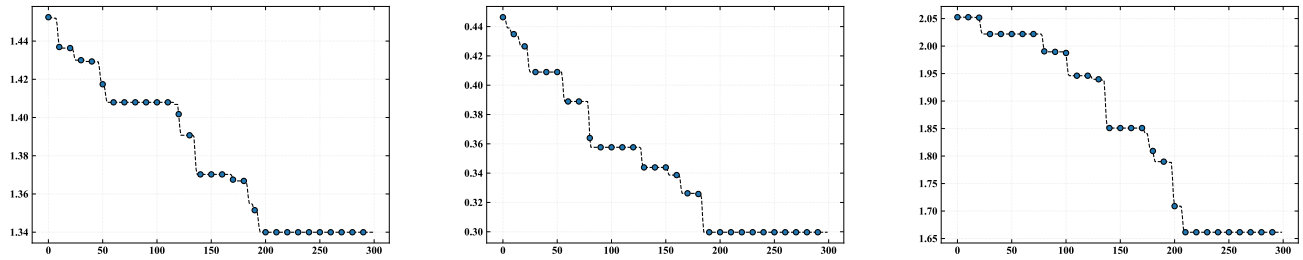
5 Result and Analysis

5.1 Comparison with Baselines

In this section, we compare the transformer-based model (ours) after knowledge selection and TAS, with the above benchmark model. The baseline results are from the corresponding paper, and all test sets and evaluation metrics are the same. The results are shown in Table 5. Among them, the performance of CLSTM in short conversation datasets (such as DailyDialog) is slightly better than CNN. However, the performance in long conversation datasets (such as MELD and EmoryNLP) is obviously inferior to that of CNN.

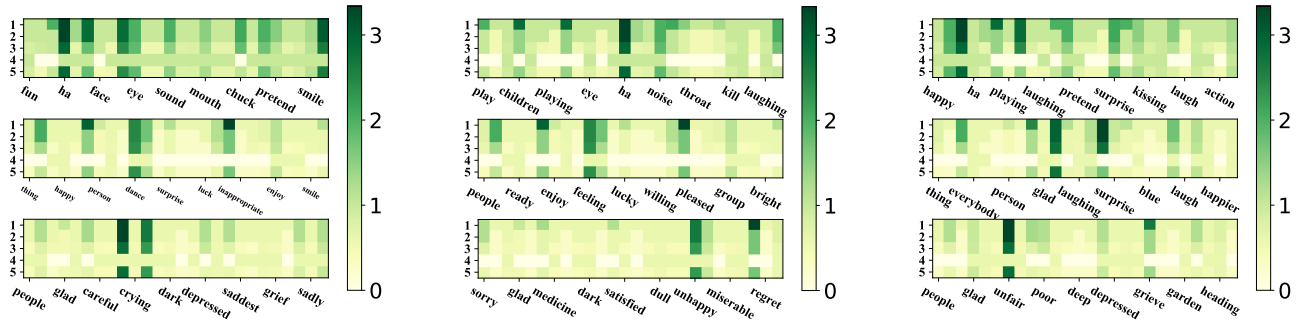
TABLE 4
The searched hyperparameters and algorithms via TAS. (Legend: "Default" represents the default parameters and algorithms combination in G&LESM.)

# Hyperparameter	# MELD	# DailyDialog	# EmoryNLP	# Default
The size of context windows (M)	6	7	6	6
The size of Glove embedding (d)	200	300	300	100
The number of hidden layers of PN	100	200	200	100
The number of heads of MSAT	4	5	4	4
Optimizer	RMSProp	Adam	Adam	Adam
Batch size	64	16	12	64
Learning rate	1.00E+04	1.00E+04	1.00E+04	1.00E+04
Loss function	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy



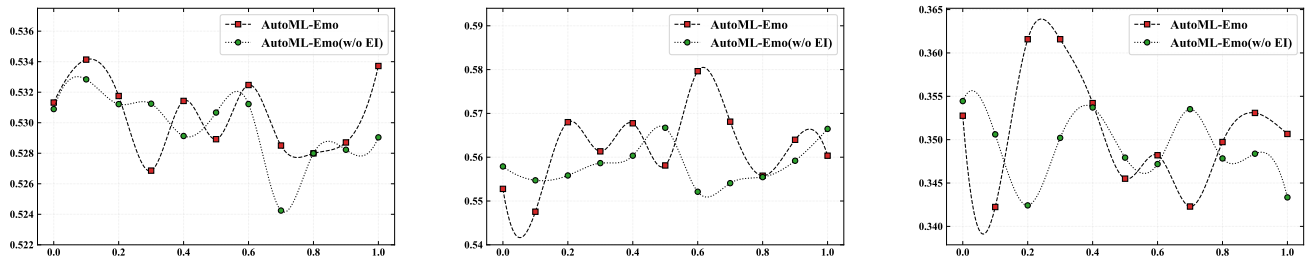
(a) MELD (b) DailyDialog (c) EmoryNLP

Fig. 3. Analysis results of LESM in MELD DailyDialog and EmoryNLP datasets. (Legend: because we use the elite operator, the curve of fitness will remain unchanged if the next generation is not as good as the current elite children.)



(a) MELD (b) DailyDialog (c) EmoryNLP

Fig. 4. The weight matrices between the words "laugh", "happy", "sad" and their related concepts. Top: weight matrix of "laugh". Middle: weight matrix of "happy". Bottom: weight matrix of "sad". (Legend: 1: the results of raw weight matrix. 2: the results after GESM w/o sentimental intensity. 3: the results after GESM. 4: the binarized results after LESM. 5: the results after LESM.)



(a) MELD (b) DailyDialog (c) EmoryNLP

Fig. 5. Analysis results of GESM in MELD DailyDialog and EmoryNLP datasets. (Legend: for showing the results clearly, we smooth the F1 score curve by interpolation technique)

In addition, the performance of DialogueRNN with attention mechanism in the long conversation dataset is still inferior to CNN. This further verifies the limitation of only using the RNNs model to capture context dependencies, which is the reason why we use the transformer-based model. Similarly, BERT_BASE and KET are also based on the transformer, but different from ours, BERT_BASE uses a bi-directional transformer to model context, which has a more powerful representation ability. However, this also makes the model parameters much more than other baselines and our model, which is extremely unfriendly to devices with limited computing power and memory. Att-NDE is performed at all times over the hidden states for different kinds of irregular time signals. The missing information in sequence data due to sampling loss, especially in the presence of long sequence, can be seamlessly compensated and attended in learning representation. Co-GAT uses a proposed co-interactive graph interaction layer where a cross-utterances connection and a cross-tasks connection are constructed and iteratively updated with each other, achieving to consider the two types of information simultaneously. HiTransformer uses speaker embedding in the model, which allows our model to capture the interaction between speakers and better understand emotional dynamics in dialog systems. Likewise, KET and HiTransformers are also not perform commonsense selection, which limits its ability to understand sentiment to some extent. As for KET, it benefits from the embedding of external knowledge riched semantics of utterances, so it performs best in all baseline methods. Unfortunately, the KET lacks an effective knowledge selection mechanism to select appropriate knowledge to reduce the negative impact of irrelevant knowledge. Additionally, the performance of KET can be improved again after an effective knowledge selection, which illustrates the great scalability of our methods. In particular, our TAS can search out the optimal combination of algorithms and hyperparameters of the transformer-based model in the internal parameter spaces, which can further improve the performance of ours and make it better than the most advanced model KET in datasets, on different sizes and domain datasets.

TABLE 5
Comparison against various baselines. (Legend: "Ours" represents a transformer-based model after knowledge selection. "Ours + TAS" represents a transformer-based model after knowledge selection and TAS. "Ours + KET" represents the KET model processed by G&LESM for knowledge selection and best values are highlighted in bold.)

Methods	# MELD	# DailyDialog	# EmoryNLP
cLSTM	49.72	49.90	26.01
CNN	55.86	49.34	32.59
BERT_BASE	56.21	53.12	33.15
DialogueRNN	56.27	50.65	31.70
Co-GAT	-	51.00	-
HiTransformer	-	-	33.04
KET	58.18	53.37	34.39
Att-NDE	56.50	-	-
Transformer-based	54.75	51.44	31.60
Ours	55.34	53.24	34.31
Improvement ↑	0.59%	1.80%	2.17%
Ours + TAS	56.89	54.20	35.37
Improvement ↑	2.14%	2.76%	3.77%
Ours + KET	**58.66**	**54.82**	**35.77**
Improvement ↑	0.48%	1.45%	1.38%

TABLE 6
The results of ablation study. (Legend: "EI" represents sentimental intensity.)

Methods	# MELD	# DailyDialog	# EmoryNLP
Ours	55.34	53.24	34.31
Ours + TAS	56.89	54.20	35.37
Ours (w/o LESM)	55.05	53.10	33.15
Ours (w/o LESM & EI)	54.93	52.31	32.42
Ours (w/o GESM & LESM)	54.75	51.44	31.60

5.2 Model Analysis

Analysis of GESM: the goal of GESM is to select suitable knowledge from the unknown external knowledge bases quickly and globally, so as to effectively reduce the scale of commonsense knowledge and pave the way for further selection. Although GESM is a rough search strategy based on grid search, it benefits from the emotion congruent effect. Thus, its selection strategy makes the selected knowledge useful in most cases. In Fig. 5, we show the accuracy and F1 score of the evaluation model on validation datasets under different Ω^* conditions without considering the sentiment intensity. Obviously, using sentiment intensity to measure the degree of emotional consistency plays a positive role. Additionally, the performance of the model is improved at first and then decreased after reaching the extreme value with the increase of the model. This is good proof of our point of view, that is: not all commonsense knowledge is very important, and the introduction of a large amount of irrelevant knowledge is bound to bring a lot of noise to data. This does not mean that the scale of knowledge should be small enough, because only a small amount of knowledge is used to enrich the semantics of the text, and the benefits are negligible. Noticeably, GESM leverages the relationship between the scale of knowledge and the performance of models from the perspective of emotional congruence.

Analysis of LESM: the goal of LESM is to select more suitable knowledge from the known knowledge accurately and locally, so as to effectively improve the quality of introduced knowledge and better adapt to the data of different sizes and domains. Unlike GESM, LESM is a search strategy based on a genetic algorithm. It only uses the fitness function (the validation loss) as the selection basis to further search the existing results accurately. Therefore, when searching the local known knowledge, it will not be affected by emotion, which makes up for the deficiency of GESM to a certain extent, that is, the limitation of positive words in a negative context. In Fig. 3, We show that with the increase in the number of generations, the fitness of the population decreases gradually. Obviously, it is necessary to use a genetic algorithm to further select the results of GESM. It is worth noting that the experimental cost of the algorithm is expensive if the knowledge is directly selected by LESM without GESM, and the search results are easy to fall into local optima. Just because of the cooperation and mutual promotion between LESM and GESM, our proposed AutoML-Emo can effectively improve the quality of external knowledge shown in Fig. 4.

Analysis of TAS: the goal of the transformer-based model is to incorporate knowledge and capture the context information in conversations. The TAS is a strategy to automatically search the optimal hyperparameters and algorithms combination in the internal parameters of the model randomly. It finally selects the combination on the validation sets to make the model perform best. If there is no TAS, the transformer-based model is constructed directly by default, such as the default transformer-based model in G & LESM. This can not show the real performance of models and will have a negative effect on the results and conclusions. However, the experimental cost of manual parameter adjustment is too high. Fortunately, this kind of extensive experiment can be effectively addressed by TAS. As shown in Table 5, the performance of the transformer-based model after TAS has been significantly improved on different sizes and domain datasets and is better than the most advanced model.

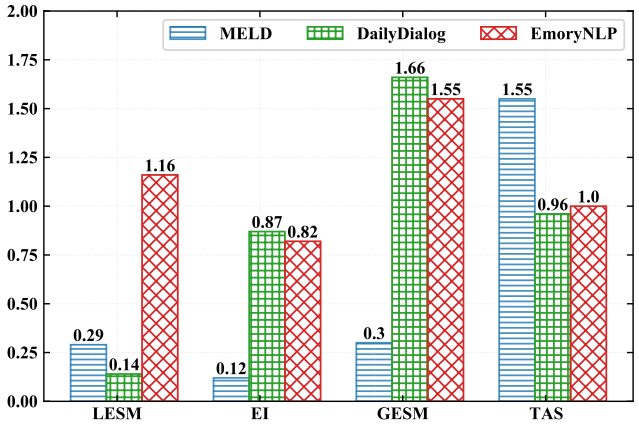


Fig. 6. Analysis results of ablation study. (Legend: the greater the lifting F1 score of different structures, the more contribution in AutoML-Emo.)

5.3 Ablation Study

In this section, we conducted ablation studies to analyze the contribution of different structures in AutoML-Emo, as shown in Table 6. Obviously, whether LESM, GESM, and TAS, they can promote the performance of the transformer-based model. However, as shown in Fig. 6, TAS has the largest contribution to the MELD dataset, which benefits from the optimal combination of hyperparameters and algorithms based on AutoML strategy. In DailyDialog and EmoryNLP, GESM has the greatest contribution, which thanks to the inspiration of the emotion congruent effect and the influence of emotional intensity. However, GESM has not achieved satisfactory results in the MELD dataset. One possible explanation is that there are a lot of satirical elements on MELD, which are difficult to understand only from the text. For example: "that's great. I'm going to enjoy it on the balcony. I can enjoy my scenery and my dessert at the same time". From the appearance, it shows that the speaker is very satisfied with his dessert and hopes to improve the experience by enjoying it on the balcony. However, careful observation of the speaker's facial expression helps us to understand the speaker's aversion to desserts, resulting in negative emotions in the process of speaking [59], [60]. The knowledge selection strategy is based on emotional congruence, in the above situation, if only from the text level, the results are often contrary to the wishes.

6 Conclusion and Future

We propose an AutoML strategy based on emotion congruent effect, which is called AutoML-Emo. On the one hand, it can effectively select the huge and complex external commonsense knowledge to improve the quality of reference knowledge; on the other hand, it can select the best hyperparameters and algorithms combination for the transform-based model. It consists of three parts: GESM, LESM, and TAS. Among them, GESM is a knowledge selection mechanism based on emotion congruent effect, which is used to quickly and globally select unknown knowledge to avoid falling into local optima. In addition, GESM also uses sentimental intensity to further describe the degree of emotional consistency. The LESM is an automatic selection mechanism based on a genetic algorithm, which is used to accurately and locally select more suitable knowledge combinations from known external knowledge. If there is no GESM and LESM research knowledge directly, the experimental cost of the algorithm is expensive, and the search results are easy to fall into local optima. On the contrary, if there is no LESM, GESM will not work when positive words appear in a negative context. At this time, LESM makes up for the shortcomings of GESM. It is because of the cooperation and mutual promotion between LESM and GESM that AutoML-Emo can achieve such an exciting effect in knowledge selection. In addition, LESM and GESM can not only promote the performance of our transformer-based model but also promote the advanced method KET, which illustrates its good scalability. As for TAS, it is similar to NAS, but its search is limited to the internal parameter spaces of the transform-based model. Its existence saves a lot of experimental costs in the optimization process of the hyperparameters and algorithms involved in models. After the TAS, the performance of the transformer-based model has been significantly improved, and it is better than the most advanced model.

In the future, we will continue to integrate word-level, utterance-level, context-level, and dialogue-level multimodal emotions to guide the knowledge selection of AutoML-Emo. In addition, the measurement of emotional consistency will not only depend on the emotional intensity obtained from the VAD emotion model. More diversified measurement methods should be considered, such as the hourglass emotion model to measure the degree of emotional consistency. Because the hourglass is a hybrid model combining discrete method and dimension method.

Acknowledgments

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62106136, 61902231), Natural Science Foundation of Guangdong Province (2019A1515010943), The Basic and Applied Basic Research of Colleges and Universities in

Guangdong Province (Special Projects in Artificial Intelligence)(2019KZDZX1030), 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D), Science and Technology Major Project of Guangdong Province (STKJ2021005), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology(2020B1212060078).

REFERENCES

[1] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022.

[2] G. Xiao, G. Tu, L. Zheng, T. Zhou, X. Li, S. H. Ahmed, and D. Jiang, "Multi-modality sentiment analysis in social internet of things based on hierarchical attentions and csattcn with mbm network," *IEEE Internet of Things Journal*, 2020.

[3] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, and D. Jiang, "Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models," *Knowledge-Based Systems*, p. 107598, 2021.

[4] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4762–4779.

[5] P. Zhong, D. Wang, and C. Miao, "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, pp. 165–176.

[6] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2470–2481.

[7] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 1033–1038.

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[9] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2122.

[10] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention." in *IJCAI*, 2018, pp. 4623–4629.

[11] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, "Towards exploiting background knowledge for building conversation systems," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2322–2332.

[12] J. Zhou, J. X. Huang, Q. V. Hu, and L. He, "Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification," *Knowledge-Based Systems*, vol. 205, p. 106292, 2020.

[13] E. Cambria, S. Poria, A. Hussain, and B. Liu, "Computational intelligence for affective computing and sentiment analysis [guest editorial]," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 16–17, 2019.

[14] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, and D. Jiang, "Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models," *Knowledge-Based Systems*, vol. 235, p. 107598, 2022.

[15] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[16] B. Chen, H. Wu, W. Mo, I. Chattopadhyay, and H. Lipson, "Autostacker: A compositional evolutionary learning system," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2018, pp. 402–409.

[17] S. Estevez-Velarde, Y. Gutiérrez, A. Montoyo, and Y. Almeida-Cruz, "Automl strategy based on grammatical evolution: A case study about knowledge discovery from text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4356–4365.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] P. Ekkekakis, *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press, 2013.

[20] J. Zhang, H. S.-H. Chung, and W.-L. Lo, "Clustering-based adaptive crossover and mutation probabilities for genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 3, pp. 326–335, 2007.

[21] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2019, pp. 527–536.

[22] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 986–995.

[23] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in *Workshops at the thirty-second aaai conference on artificial intelligence*, 2018.

[24] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, 2020.

[25] G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria, "Context-and sentiment-aware networks for emotion recognition in conversation," *IEEE Transactions on Artificial Intelligence*, 2022.

[26] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[27] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[28] L. Devillers, I. Vasilescu, and L. Lamel, "Annotation and detection of emotion in a task-oriented human-human dialog corpus," in *proceedings of ISLE Workshop*, 2002.

[29] D.-A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Transactions on Affective Computing*, 2018.

[30] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system," *IEEE Annals of the History of Computing*, no. 01, pp. 1–1, 2020.

[31] A. Hussain, E. Cambria, S. Poria, A. Y. Hawalah, and F. Herrera, "Information fusion for affective computing and sentiment analysis," pp. 97–98, 2021.

[32] Y. Ma, K. L. Nguyen, F. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[33] C. Liu, S. Wu, D. Jiang, Z. Yu, and H.-S. Wong, "View-aware collaborative learning for survival prediction and subgroup identification," *IEEE Transactions on Biomedical Engineering*, 2022.

[34] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017, pp. 873–883.

[35] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020.

[36] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

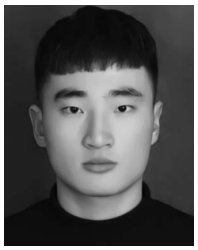
[37] E. Cambria, Q. Liu, S. Decherchi, F. Xing, , and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *LREC*, 2022, pp. 3829–3839.

[38] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *AAAI*, 2018, pp. 4970–4977.

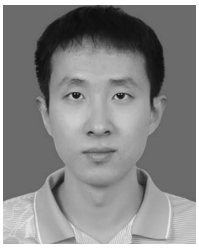
[39] D. Zhang, X. Chen, S. Xu, and B. Xu, "Knowledge aware emotion recognition in textual conversations via multi-task incremental

transformer," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4429–4440.
[40] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.
[41] V. Lopes, A. Gaspar, L. A. Alexandre, and J. Cordeiro, "An automl-based approach to multimodal image sentiment analysis," *arXiv preprint arXiv:2102.08092*, 2021.
[42] M. T. AL-Sharuee, F. Liu, and M. Pratama, "Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison," *Data & Knowledge Engineering*, vol. 115, pp. 194–213, 2018.
[43] Z. Chen, N. Ma, and B. Liu, "Lifelong learning for sentiment classification," *arXiv preprint arXiv:1801.02808*, 2018.
[44] S.-Y. Su, P.-C. Yuan, and Y.-N. Chen, "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, Volume 1 (Long Papers)*, 2018, pp. 2133–2142.
[45] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
[46] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.
[47] T. Elsken, J. H. Metzen, F. Hutter *et al.*, "Neural architecture search: A survey." *J. Mach. Learn. Res.*, vol. 20, no. 55, pp. 1–21, 2019.
[48] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.
[49] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.
[50] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[51] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, pp. 309–317, 2019.
[52] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 174–184.
[53] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The Hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.
[54] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://www.aclweb.org/anthology/D14-1181
[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[56] Y.-H. Chen and J.-T. Chien, "Continuous-time attention for sequential learning," in *Proc. of AAAI Conference on Aritificial Intelligence*, 2021.
[57] L. Qin, Z. Li, W. Che, M. Ni, and T. Liu, "Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification," *arXiv preprint arXiv:2012.13260*, 2020.
[58] Q. Li, C. Wu, Z. Wang, and K. Zheng, "Hierarchical transformer network for utterance-level emotion recognition," *Applied Sciences*, vol. 10, no. 13, p. 4447, 2020.
[59] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.
[60] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

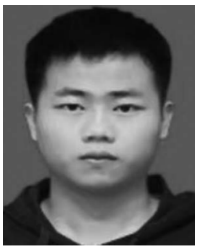
Dazhi Jiang received his BA in Computer Science from the China University of Geoscience (Wuhan) in 2004. He obtained his PhD from the State Key Laboratory of Software Engineering, Wuhan University, China in 2009. Since then, he has been with the Department of Computer Science, Shantou University, China where he was a Professor. His research interests include affective computing, deep learning, data mining and applications of artificial intelligence.



Runguo Wei is currently pursuing the master's degree with the Department of Computer Science at Shantou University, China. His current research focuses on affective computing and machine learning, etc.



Jintao Wen is a graduate student of the Department of Computer Science at Shantou University. His current research interests includeÂăaffective computing and deep learning.



Geng Tu is a graduate student of the Department of Computer Science at Shantou University. His current research interests includeÂăaffective computing and deep learning.

Erik Cambria (Fellow, IEEE) received the Ph.D. degree in computing science and mathematics from the University of Stirling, UK, in 2012, following the completion of an EPSRC project in collaboration with MIT Media Lab, USA. He is currently a Full Professor at Nanyang Technological University, Singapore. His research focuses on neurosymbolic AI for explainable natural language processing in domains, such as sentiment analysis, dialogue systems, and financial forecasting. Dr. Cambria is recipient of several awards, including the 2018 "AI's 10 to Watch" and the 2019 "IEEE Outstanding Early Career Award", and was featured in Forbes as "One of the 5 People Building our AI Future".