IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 September 2023, accepted 13 October 2023, date of publication 19 October 2023, date of current version 27 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325893

RESEARCH ARTICLE

Directed Acyclic Graphs With Prototypical Networks for Few-Shot Emotion Recognition in Conversation

YUJIN KANG AND YOON-SIK CHO
Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Yoon-Sik Cho (yoonsik@cau.ac.kr)

This work was supported in part by the Chung-Ang University Research Scholarship Grants, in 2022; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (MSIT) through the Development of Emotion Recognition/Generation-Based Interacting Edge Device Technology for Mental Health Care under Grant 2022-0-00331; and in part by the Artificial Intelligence Graduate School Program (Chung-Ang University) under Grant 2021-0-01341.

ABSTRACT Emotion recognition in conversation is the task of recognizing the emotion in each utterance of a conversation, and it is an active field of study with various applications. Many studies have been conducted in well-designed settings in which all of the emotion labels in the training set are available. However, a rich dataset with emotion labels for all utterances is rare. In this study, we address this problem by resorting to few-shot learning. Specifically, we propose Directed Acyclic Graph (DAG)-based prototypical networks, namely **ProtoDAG**, to better capture contextual information in conversations, which in turn leads to accurate emotion predictions. In our model, the DAG layers are tailored into prototypical networks, which is learned end-to-end. Our experiments with popular benchmark datasets demonstrate that our model achieves state-of-the-art results outperforming the existing few-shot learning model by a significant margin and is even competitive with fully supervised baseline models for emotion recognition in conversation.

INDEX TERMS Directed acyclic graph, emotion recognition in conversation, episodic learning, few-shot learning, prototypical network.

I. INTRODUCTION

Motivated by the success of personal assistance services, Emotion Recognition in Conversation (ERC) has become an active research field. ERC is the task of predicting the emotions behind utterances in a conversation. For this prediction task, many researchers have resorted to supervised learning [1], [2], [3], [4], [5], which generally requires accurate high-volume labeling. However, labeling each utterance with its emotion requires extensive human labor. Annotating the emotion for each utterance requires selecting one from six or seven emotions [6], which makes the task challenging. One circumvention method utilizes synthetic conversations such as scripts for TV shows [7], [8] or acting [9].

However, even with these alternatives, the supervised approach to ERC is difficult because it inherently requires full pairs of utterances and emotions. Few-Shot Learning (FSL) tackles this problem by using few samples, and it can reduce the data gathering requirement. Episodic learning [10] is a popular strategy for training FSL methods, which learns over multiple *episodes*. The episodes are sampled from a task family, and these methods employ a general purpose learning algorithm that can generalize across tasks. FSL models with episodic learning are robust to data generalization, and their performance is less affected by dataset size. In few-shot scenarios, prototypical networks [11] use this training strategy combined with neural networks for non-linear embeddings. Prototypical networks learn non-linear embedding space in which classification can be performed using distances to prototype representations. With the success of prototypical networks in many applications, Guibon et al. [12] first applied prototypical networks in

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

ERC. While their work is meaningful as the first application of FSL in ERC, the performance was not as significant as that of the full supervised learning approach.

To this end, we revisit episodic learning with prototypical networks for ERC. Specifically, we focus on strengthening the context embeddings in conversations. We also draw inspiration from the Directed Acyclic Graph Emotion Recognition in Conversation (DAG-ERC) model [3]. DAG-ERC proposes an intuitive method that combines the Recurrent Neural Network (RNN) and Graph Neural Network (GNN) to understand the context flow. While the DAG layer in DAG-ERC [3] is effective, we raise doubts that their emotion prediction ignores the neighboring relation embeddings when all the node embeddings are flattened in their softmax layer. We hypothesize when the classification layer is replaced by other distance-based metric in the prediction module, two modules: the emotion prediction and the GNN in DAG can achieve pure embeddings, where the two modules align more towards the same direction. Thus, we extend DAG-ERC into prototypical network for the context encoding in an end-to-end manner. We propose a DAG-based Prototypical Networks, which leverages few-shot learning to address the data insufficiency and DAG to capture the unidirectional nature of dialogues and provide rich contextual understanding. Our contribution is three-fold.

- We contribute by proposing directed acyclic graph-based prototypical networks, which we name as **ProtoDAG**, to address ERC in a few-shot scenario. To the best of our knowledge, this study is the first in few-shot ERC based on GNNs.
- We conduct experiments on two benchmark datasets: DailyDialog and IEMOCAP and achieved state-of-the-art results with a significant margin.
- Moreover, our FSL model now performs comparable with supervised learning models with full labels.

II. RELATED WORK

A. RNN-BASED MODELS

One of the earliest works on context modeling for ERC involves long short-term memory (LSTM) using multimodal data [13]. This work was followed by ICON [14] and CMN [15] both of which applied the Gated Recurrent Unit (GRU) [16] to multimodal data. These three models used multimodal data instead of focusing only on textual information in conversations. HiGRU [17] applied two levels of bidirectional GRUs; the lower-level GRU is used for utterance embedding, and the upper-level GRU is used to capture a sequential relationship in context. DialogueRNN [1] used three types of GRUs to express the significant factors that influence the emotion behind utterances in a conversation. The most recent work based on RNNs is COSMIC [2], which extends DialogueRNN by applying commonsense knowledge to it. However, these RNN variant models exhibit drawbacks such as the inability to capture long-term dependencies or speaker relationships beyond sequences as the models have vanishing gradient problem and limitation of fixed-length contexts. While the first drawback can be reduced by stacking multiple RNN layers, the model is still limited by the amount of information that can be input into the model.

B. GNN-BASED MODEL

To resolve the limitations of RNN-based models, GNN-based models have been proposed. DialogGCN [18] uses the Graph Convolutional Network (GCN) to leverage the self and inter-speaker dependency of the interlocutors to model the conversational context. This method is known to resolve the long-term information propagation issues of the RNN. RGAT [19] further strengthened the relationship between utterances by adding positional encoding with a self-attention mechanism. KET [20] added external knowledge to Transformer, which is similar in spirit to COSMIC with an RNN. DialogXL [21] used XLNet, which is an extension of Transformer-XL, to improve understanding of the context. S+PAGE [4] sought to understand the context using both Transformer and the R-GCN [22], and it is currently the state-of-the-art model on one of the benchmark datasets. However, graph-based models still suffer from capturing dependencies between distant utterances or sequential information. To remedy the shortcomings of both the RNN and GNN, DAG-ERC [3] combined the GRU and DAGs. Most recently, Yang et al. [5] proposed a hybrid curriculum learning framework for the ERC task, and they also extended previous models including DAG-ERC with hybrid curriculum learning. While the aforementioned methods achieved promising results in ERC, all of these models require a sufficient volume of labeled data. This limitation leads us to review FSL for ERC.

C. FEW-SHOT LEARNING

Several papers [23], [24], [25] have noted the difficulty of constructing a large dataset for learning, and they therefore addressed the need for few-shot learning. Few-shot learning aims to learn effectively with only a small number of samples per class. FSL methods use episodic learning [10] to simulate an environment in which labeled data is scarce, and metric-based learning is a method of representative FSL. As the episodic learning constructs multiple episodes, each comprising a small subset of the training data, the model with episodic learning can quickly learn and generalize from a limited yet diverse dataset. Metric-based learning learns the similarity of data by measuring the distance between data. The Siamese network [26] is a model that compares the distance between the results of two networks using weighted L1. Triplet networks [27] adjust the distance of data based on the baseline data with Siamese networks. Matching Networks [28] use the Convolutional Neural Network (CNN) and Bidirectional LSTM (BiLSTM) as embedding functions and cosine similarity to calculate the distance between the data. Prototypical networks [11] compute the distance between query points and prototypes and assign query points to their closest cluster with its associated class label. Here, a distance metric such as the Euclidean distance is used, and the mean of the embedded support examples for each class becomes the prototype. Sung et al. [24] further proposed a *learns to learn* framework that learns a deep distance metric in an end-to-end fashion.

Prototypical networks have shown promising results in FSL, and they have been mostly focused on the image domain. Recently, prototypical networks have also been used in various Natural Language Processing (NLP) tasks [29], [30], [31], [32]. For ERC, ProtoSeq [12] is currently the only approach that uses FSL. In ProtoSeq, the output from the context encoder is fed into a multi-layer perceptron (MLP) for prototype creation, and emotions are predicted through prototypical networks and conditional random fields. However, ProtoSeq neglects important features in conversations such as information flow between the long-distance conversation background and nearby context.

III. METHODOLOGY

A. TASK DEFINITION

We consider a conversation with a sequence of utterances $C = (u_1, u_2, \cdots, u_n)$, where n is the number of utterances. Each utterance u_i consists of word tokens $u_i = \{w_{i1}, w_{i2}, \cdots, w_{im}\}$, where m represents the number of tokens. Each utterance u_i corresponds to its speaker and its emotion label y_i. All conversations consist of up to two speakers unless otherwise specified. The task aims to determine the emotion behind each utterance in a conversation.

B. EPISODIC APPROACH

For supervised learning, a large volume of labeled data is used as input. However, annotating emotion labels for every utterance is expensive in terms of both time and resources, which provides a motivation for using FSL. We approach FSL with episodic learning [10], which is similar to previous work, including ProtoSeq [12]. The episodic approach limits the number of data per class to N, constituting several distinct datasets within one epoch. Although the size of the dataset created with episodic learning is small, the class of the original dataset can be observed, and therefore, the characteristics of each class can be learned.

C. MODEL

We addressed ERC under challenging settings, where only a small amount of data with emotion labels is provided, and we approached the problem with few-shot learning based on prototypical networks [11]. Prototypical networks have a simple inductive bias, so they can make predictions even in environments with limited data. Understanding utterances as well as the flow of utterances in the ERC task is crucial; it is an important factor in terms of improving the predictive performance. The importance of context also applies to FSL settings, and we can build models that focus on more context-driven embeddings using these simple few-shot learning techniques. The overall procedure (per each episode) of the model is provided in Algorithm 1. In each epoch, the Algorithm 1's process is repeated as many times as the number of episodes. We divide our model into five stages, and describe each of the stage in the following.

Algorithm 1 This Algorithm Demonstrates How the Utterances in One Conversation Are Predicted Using Prototypical Networks. Sampling($\mathcal{D}$,N_s) Represents the Data, Which Is Randomly Chosen From Dataset $\mathcal{D}$ as Many as the Number of Shots N_s and f Denotes the Utterance Feature Extracting Function.

```
Input:
$N_s, N_q, N_w$: the number of shots, queries, ways(classes);
$\mathcal{D} = (C_1, C_2, \ldots, C_N)$: whole dataset and each conversation
$C_i = \{(u_{i1}, y_{i1}), \ldots, (u_{i|C_i|}, y_{i|C_i|})\}$
Output: A prediction $P$ for utterances in conversation
for $k$ in $\{1, \ldots N_w\}$ do
    $Support_k \leftarrow$ sampling($\mathcal{D}, N_s$)
    $Query_k \leftarrow$ sampling($\mathcal{D} - Support_k, N_q$)
end for
for $k$ in $\{1, \ldots N_w\}$ do
    $S \leftarrow \varnothing$
    for $C_s$ in $Support_k$ do
        $T_s \leftarrow \varnothing$
        for $(u_s, y_s)$ in $C_s$ do
            $T_s \leftarrow T_s + f(u_s)$
        end for
        $S \leftarrow S \cup$ DAG-ERC Layers($T_s$)
    end for

    $c_k \leftarrow \frac{1}{N_s} \sum_{\mathbf{s_i} \in S} \mathbf{s_i}$
end for
$Q \leftarrow \varnothing$
for $C_q$ in $Query_{k=\{1,\ldots,N_w\}}$ do
    $T_q \leftarrow \varnothing$
    for $(u_q, y_q)$ in $C_q$ do
        $T_q \leftarrow T_q + f(u_q)$
    end for
    $Q \leftarrow Q \cup$ DAG-ERC Layers($T_q$)
end for
$P \leftarrow \varnothing$
for $\mathbf{q_i}$ in $Q$ do
    for k in $\{1, \ldots N_w\}$ do
        $P \leftarrow P \cup \arg\min_k L_2(\mathbf{c_k}, \mathbf{q_i})$
    end for
end for
return $P$
```

1) THE SUPPORT AND QUERY SETS

The support and query sets are major components of FSL. For each class (or emotion in ERC), the support set is constructed by collecting N_s conversations with a sequence

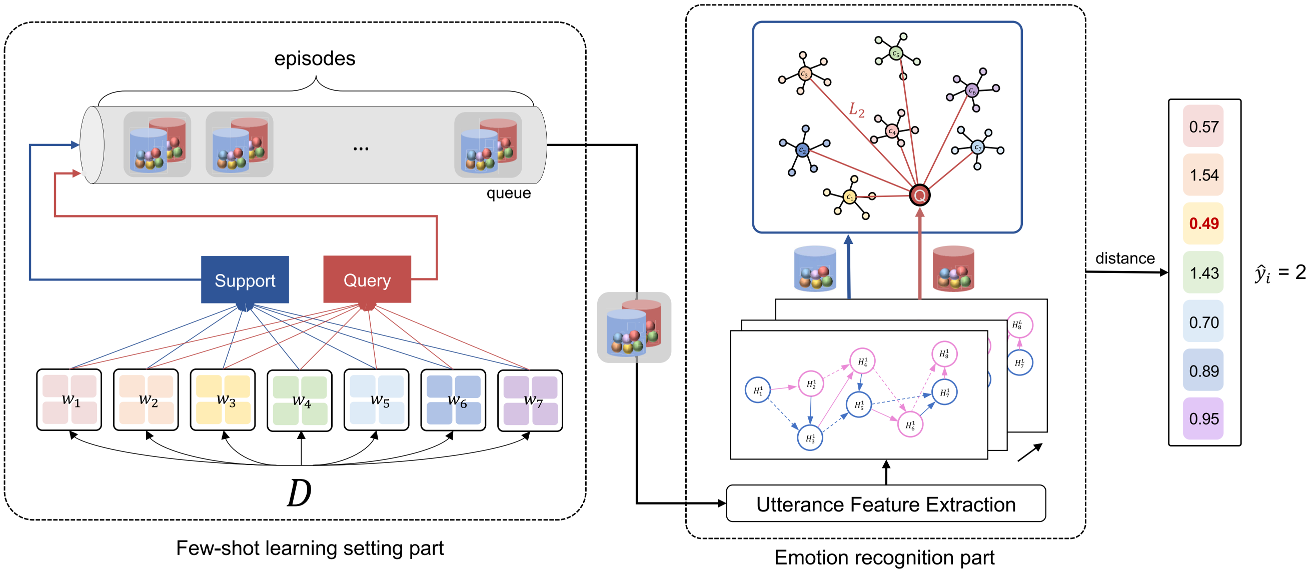


FIGURE 1. **ProtoDAG Framework using a 7-way 5-shot 10-query configuration. D represents the dataset, w indicates the class (way), and $\hat{y}_i$ represents the prediction for the i-th utterance. The actual prediction in ProtoDAG is performed over a query set, which we simplify to a single utterance in this Figure. The emotion recognition is repeated as much as episodes, and an epoch ends when the queue becomes empty.**

of utterances and the corresponding emotion labels from the entire dataset. We repeat this support set construction process for all N_w classes. In the context of FSL, N_s is the number of *shots*, and N_w is the number of *ways*. Overall, the support set consists of $N_s * N_w$ conversations. In the same way, the query set consists of $N_q * N_w$ conversations. This process is shown in line 1-4 in Algorithm 1. While normal supervision learning trains data by dividing it into batch sizes, ProtoDAG trains models using support and query sets. In our dataset, DailyDialog, the task has a 7-way 5-shot 10-query configuration. The size of the DailyDialog support set is 35 (#shots * #ways). When using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, the task is 6-way 1-shot 5-query, and the size of the support set is 6. The number of utterances constituting one conversation and the number of sentences constituting one utterance is flexible. To enable a direct comparison with the previous baseline in few-shot emotion recognition in conversation, we set up a 7-way 5-shot 10-query configuration for DailyDialog dataset.

2) UTTERANCE FEATURE EXTRACTION

We use RoBERTa-Large [33], a Transformer-based pre-trained language model for utterance embeddings. [CLS] token is prepended to the beginning of each utterance to imply the meaning of the entire utterance as shown in Equation 1, and the embedding of the last layer of [CLS] is used as an utterance feature [3]. The size of the hidden vectors in the feature extractor is 300, and the feature size is 1024.

$$u_i = f(\{[\mathrm{CLS}], w_{i1}, w_{i2}, \ldots, w_{in}\}), \tag{1}$$

where function f denotes the feature extractor from RoBERTa-Large, and n is the the number of tokens in u_i.

3) DIRECTED ACYCLIC GRAPH NEURAL NETWORK FOR CONTEXT EMBEDDING

In the process of using DAGs for emotions in conversation, each conversation enters the embedding stage. Using the DAG-ERC layers, the embedding of the i-th utterance of the l-th layer is calculated by the neighboring nodes $\mathcal{S}_i$ of u_i^l as well as the hidden state of the previous layer. The neighboring nodes $\mathcal{S}_i$ of u_i consist of all utterances that exist before the current node u_i and share the same speaker as the current node.

First, the information from the peripheral node is calculated through the relationship between the current node and the surrounding nodes. Parameter $W_{r_{ij}}$ indicates whether the speaker of u_i and u_j is the same, and aggregated information M_i^l of u_i is calculated using the attention weight α_{ij}, relation aware parameter $W_{r_{ij}}$ and hidden state H_j^l of the peripheral node in the current layer. Attention weight α_{ij} is calculated by concatenating the hidden state in the previous layer of the current node u_i and the hidden state in the current layer of the nodes, which is the predecessor of u_i.

$$M_i^l = \sum_{j \in \mathcal{S}_i} \alpha_{ij} W_{r_{ij}}^l H_j^l, \tag{2}$$

where $W_{rij}^l \in \{W_0^l, W_1^l\}$ are trainable parameters for the relation-aware transformation [22] and $\mathcal{S}_i$ is a set of peripheral nodes that have been uttered before u_i. Shen et al. have claimed only using the nodal information unit is not enough for ERC, and have introduced the contextual information unit. The nodal information unit (GRU_H) uses H_i^{l-1}as its input and M_i^l as its hidden state; the contextual information unit (GRU_M) uses M_i^l as its input and H_i^{l-1} as its hidden state. The nodal information unit focuses on the node information propagating from the past layer to the current layer. The contextual information unit allows better understanding of the context by using M_i^l as input and the H_i^{l-1} as its hidden state.

In Equation 3, we show the process of computing one node. The node is calculated by inputting the previous layer value (H_i^{l-1}) as input and the aggregated value (M_i^l) as a state. In the GRU, M_i^l is used as a state to adjust the propagation of the hidden state of u_i. During the ERC task, the most vital clue as to the emotions is context. We try to focus on understanding the flow of information by using M_i^l as input and the hidden state of the node's previous layer as hidden state by using one more GRU to recognize emotions using the context. Finally, two GRUs are used to embed both nodes and contexts.

$$H_i^l = \underbrace{\mathrm{GRU}_H^l(H_i^{l-1}, M_i^l)}_{\text{nodal}} + \underbrace{\mathrm{GRU}_M^l(M_i^l, H_i^{l-1})}_{\text{contextual}}. \tag{3}$$

When the embedding process is complete, the output is input into an MLP. The MLP is consists of two sets of fully connected layers that use dropout and ReLU, and then, the MLP goes through an additional fully connected layer with dropout.

4) PROTOTYPES CREATION

To create prototypes, we followed the process published in [11]. We obtain prototypes from the mean vectors of the embedded support points for each class. Examples from the support set of embedded utterances passed through the MLP form prototypes. This process can also be seen in Line 15 of Algorithm 1. Emotion prototypes are computed based on the embedding values of the support set after the embedding process of the support set (Line 5-13) is completed.

$$c_k = \frac{1}{N_s} \sum_{y_i=k,(u_i,y_i)} \mathbf{s_i}, \tag{4}$$

where N_s is the number of shots, k is class, and $\mathbf{s_i}$ is the embedding of u_i, which go through both the utterance and context embedding steps.

5) CLASSIFICATION

To perform classification after prototype creation using the support set, the Euclidean distance between the embedded value of the query set and each prototype is computed. The prototype with the minimum distance is selected, and its corresponding class label is then assigned to the query for the prediction. This is depicted as Line 28 in Algorithm 1.

D. DAG-ERC AS A BACKBONE

DAG-ERC [3] is a model that combines the advantages of conventional graph-based natural models and recurrence-based natural models to express the structure of a conversation using the GNN. Dialogue proceeds in one direction, and future utterances do not affect the past. DAG-ERC expresses the characteristics of these conversations as a DAG. DAG-ERC also divides the graph edges into remote and local information according to which information in the conversation is essential for understanding the meaning of the utterances.

TABLE 1. Statistics of the datasets. # Conv and # Uttr refer to the number of conversations and utterances, respectively. Max, Min, and Avg represent the maximum, minimum, and average values of the number of utterances in each conversation. Eval means Evaluation Metric.

Dataset	DailyDialog			IEMOCAP		
	train	val	test	train	val	test
# Conv	11118	1000	1000	120		31
# Uttr	87170	8069	7740	5810		1623
Max	35			167		
Min	2			24		
Avg	7.85			66.3		
classes	7			6		
Eval	Micro-F1			Weighted-F1		

TABLE 2. Percentage of categorical labels in DailyDialog dataset.

Label	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Num	1022	353	74	12885	1150	1823	85572
Percentage(%)	0.99	0.34	0.07	12.52	1.11	1.77	83.17

While we also rely on DAG, the DAG-ERC layers have been reformulated into prototypical network. Algorithm 1 shows how DAG-ERC layers have been incorporated into episodic learning. Our DAG-ERC layers are not only trained over small number of samples compared to that of DAG-ERC but, more importantly it is trained with totally different approach. In DAG-ERC, embedding is learned by optimizing the prediction probability according to its emotion labels, where the node embeddings from GNN are simply collected for the sigmoid input. Hence, the DAG relation from GNN could be smoothed in end-to-end training with the sigmoid prediction. Whereas, our model exploits the distance based optimization. Hence in our model, the embeddings with same emotions are close each; whereas the embeddings with different emotions are far apart. The distance based optimization overcomes the limitation of DAG-ERC [3] which uses the softmax for emotion prediction.

IV. EXPERIMENTAL SETTINGS

A. DATA

Motivated by the introduction of AI conversational systems, such as chatbots in healthcare, and customer services, emotion recognition in conversion (ERC) has attracted increasing attention in a dyadic conversation setting. We evaluated our model on two dyadic ERC Datasets: DailyDialog and IEMOCAP. Table 1 shows the statistics for both DailyDialog and IEMOCAP, and the details of datasets are also described in this section.

a: DAILYDIALOG

DailyDialog [34] is a manually labeled multi-turn open-domain dialogue dataset. The conversation is manually labeled with communication intention and emotion information. Due to its rich emotion labels, this dataset is favored by researchers in ERC. Each utterance in the

TABLE 3. Percentage of categorical labels in IEMIOCAP dataset.

Label	Happy	Sad	Neutral	Anger	Excited	Frustrate
Num	648	1084	1708	1103	1041	1849
Percentage(%)	9	15	23	15	14	25

dataset is labeled according to seven emotion categories, including anger, disgust, fear, happiness, sadness, surprise, and neutral. 83.1% of all utterances are labeled as neutral, while the remaining 16.9% of the data is labeled with one of Ekman's [6] six emotions. In terms of our evaluation metrics, the DailyDialog dataset has extremely unbalanced labels. Following previous works' chosen metrics [2], [3], [4], [5], [18], [19], [20], we chose the micro F1-score, which calculates a score that excludes the majority emotion class in DailyDialog. The neutral makes it challenging to evaluate the model's performance in distinguishing between other emotions; therefore, we exclude neutral from the dataset. Even after excluding neutral, the dataset still exhibits a significant imbalance in favor of the happy class. Therefore, we have chosen the F1-micro as the evaluation metric, which is particularly useful when dealing with severe class imbalances in the dataset.

b: IEMOCAP

IEMOCAP [9] is a multimodal ERC dataset containing video, speech, motion capture of faces, and text transcriptions. As the dataset's name implies, it was collected during dyadic sessions in which two actors performed improvisations or scripted scenarios. Originally, the dataset contained both categorical and continuous emotional descriptors. In this study, we used the categorical emotion labels and text transcriptions for consistency with the settings in DailyDialog. The original IEMOCAP provides only training and testing data, so to apply it to our model, the last 20 conversations in the training set were used as validation sets as in DialogXL [21]. The IEMOCAP dataset, while relatively more balanced compared to the DailyDialog dataset, still exhibits an imbalanced distribution, as indicated in Table 3. Therefore, to evaluate the model's performance more fairly, we have chosen the weighted F1 score as the evaluation metric.

B. TRAINING SETUP

Our code is implemented in Python using Pytorch. `ProtoDAG` is trained and tested on a single Nvidia RTX A6000 Tensor Core GPU. The hyper-parameters used in for two datasets are shown in Table 4. For fair comparison to existing FSL in ERC, we followed the settings of DailyDialog in [12]. In DailyDialog, We set the number of training, validation, and testing episodes to 100, 100, and 1000. [1] On the other hand, the IEMOCAP has a significantly smaller number of conversations than the DailyDialog; we set the number of training, validation, and testing episodes to 10.

TABLE 4. Hyper-parameters assignment for two datasets.

Hyperparameter	Dataset	
	DailyDialog	IEMOCAP
learning rate	2e-5	5e-3
dropout	0.3	0.2
# GNN layers	3	4
epoch	10000	1000
# episodes	100/100/1000	10/10/10
patience	100	100
clip gradient	None	5.0
learning rate optimizer	Adam [35]	
embedding dimension	300	
ω	1	

The total number of epochs is set to 1000 on IEMOCAP, 10000 on DailyDialog, and if the performance did not improve over 100 consecutive epochs, the learning was terminated. The DAG-ERC layers in our model have been tailored from DAG-ERC [3] to prototypical network. As we also want to compare our model to existing supervised learning approaches including DAG-ERC, we follow the published settings in [3] for each dataset. Therefore, we maintained the learning rate at 2e-5 for DailyDialog and 5e-3 for IEMOCAP. The dropout rate was set to 0.3 for DailyDialog and 0.2 for IEMOCAP. The number of GNN layers also followed DAG-ERC's setting, and it was set to 3 and 4 for DailyDialog and IEMOCAP, respectively. These numbers of layers were optimized by Shen et al. through a hyper-parameter search using their validation sets. In DAG-ERC, stacking more GNN layers allows information to be received from a remote utterance. However, the trade-off of increasing the number of layers is over-smoothing, which leads to performance degradation and should be optimized as in [3]. We also followed other published hyperparameters such as the embedding dimension as well as the cut-off point of remote and local information for both datasets; the embedding dimension was set to 300, and ω was set to 1 considering only the most recent utterance.

C. BASELINES

Here, we review the baseline supervised methods that have been applied to ERC. We then describe some few-shot approaches that can be directly compared to our model.

1) SUPERVISED LEARNING MODELS

- `DialogueRNN` [1]: A model uses three GRUs to track the speaker, context, and emotion in an utterance to recognize emotions in conversations. This work was one of the earliest work on ERC to use the RNN.
- `COSMIC` [2]: They originally incorporates commonsense knowledge to design better contextual representations for ERC. Following [3], we utilize COSMIC without incorporating external knowledge, thereby

[1] According to Guibon et al. , their settings follow [30].

TABLE 5. This table compares the results obtained using both supervised learning and few-shot learning. Supervised learning methods approach feature extraction the same way. Few-shot learning methods have unique feature extraction methods. The few-shot methods employed a 7-way 5-shot 10-query configurationn for DailyDialog and a 6-way 1-shot 5-query configuration for IEMOCAP. The main evaluation metric for supervised learning is F1 (micro) on DailyDialog, and F1(weighted) on IEMOCAP.

Method	IEMOCAP	DailyDialog
Supervised learning (full label)		
COSMIC	0.6305	0.5616
DialogueRNN	0.6476	0.5732
DAG-ERC	0.6803	0.5933
SPCL	0.6974	-
S+PAGE	0.6872	0.6407
Few-shot learning		
Proto	0.3146	0.2141
ProtoSeq	0.3781	0.3181
ProtoDAG (ours)	**0.6316**	**0.5283**

eliminating listener-specific and speaker-specific commonsense.

- DAG-ERC [3]: This model expresses the characteristics of conversation using DAGs, and the model structure was designed using the characteristics of both RNN and GNN.
- S+PAGE [4]: This model is the state-of-the-art model in the DailyDialog dataset with full labels. It leverages Transformer and the Relational Graph Convolution Network(R-GCN) on the interactions between self and inter-speakers.
- SPCL [36]: Song et al. combine the supervised contrastive learning loss and prototypical network for alleviating data imbalance in ERC dataset. It is currently the best performing model for the IEMOCAP dataset with full labels.

2) FEW-SHOT LEARNING MODELS

- Proto [11]: Proto was introduced as a baseline in [12]. It is based on the original prototypical networks [11], where the emotion labels are predicted using the Euclidean distance to class prototypes.
- ProtoSeq [12]: It is the first few-shot learning model for ERC. After encoding the context using the CNN-BiLSTM, the model predicts emotion using a prototypical networks.

V. RESULTS

The overall results are provided in Tables 5, where we compare the results of ProtoDAG to the baselines. In table, our model outperforms the existing few-shot approaches by a significant margin. Moreover, our model achieves nearly the same results as the fully-supervised models. We further discuss our results by comparing them to the fully-supervised learning baselines as well as the few-shot learning baselines.

TABLE 6. ProtoDAG performance on two datasets with varying settings on number of shots.

IEMOCAP			DailyDialog		
	shot	F1-weighted		shot	F1-micro
5-query 6-way	1	0.6316	10-query 7-way	1	0.3881
	3	0.6499		5	0.5283
	5	0.6545		10	0.5556

A. COMPARISON TO FEW-SHOT LEARNING APPROACHES

The DailyDialog dataset contains seven emotions, where we set N_W to 7. Following the same episodic composition as that published in [12], we performed few-shot learning with a 5-shot 7-way 10-query configuration.

Proto is a model that focuses on the original prototypical networks; it does not consider context and applies only embeddings to utterances. This limitation lead to a low performance of 21.41%. ProtoSeq exhibited a performance of 31.81%, which is better by 10% when compared to the Proto model. This improvement is due to the addition of the contextual embedding. After encoding the utterances with the CNN, ProtoSeq uses BiLSTM as the context encoder. Our model achieves 52.83% on DailyDialog. Rather than simply understanding the context using BiLSTM, our model understands the context more fully by using a DAG in prototypical network, which is appropriate for expressing conversational characteristics. For IEMOCAP dataset, we perform FSL with a 6-way, 1-shot 5-query configuration. Since Proto does not attempt to understand the context in conversations, it obtained a low result of 31.46%. ProtoSeq enhanced the performance from 31.46% to 37.81%. ProtoDAG outperform the ProtoSeq with significant margin from 37.81 to 63.16%, where the improvement is more remarkable than that of DailyDialog. We believe this is mainly due to the characteristic of the datasets. While IEMOCAP has limited number of conversations, all of the conversations have rich utterances. ProtoSeq relies on BiLSTM which has limitations in handling long sequences, where as the DAG in ProtoDAG better captures the rich context even in long conversations.

B. COMPARISON TO SUPERVISED LEARNING APPROACHES

On both datasets, ProtoDAG achieves remarkable results, which are even competitive with a fully supervised baseline models for emotion recognition in conversation. When compared with COSMIC [2], ProtoDAG outperforms by 0.11 on IEMOCAP, and performs below only by 3.33% on DailyDialog. Comparing to the best performing supervised learning models, they are below 11.24% and 6.58% for DailyDialog and IEMOCAP respectively. On DialyDialog, S+PAGE [4] is the best-performing supervised learning model based on speaker and position-aware conversation graph, which focuses more on self contextual feature than other approaches.

TABLE 7. ProtoDAG performance on two datasets with varying settings on number of episodes. Test Episode is fixed to 10 in IEMOCAP, 1000 in DailyDialog. Each row uses the fixed number of shots, queries.

IEMOCAP			DailyDialog		
Train	Val	F1-weighted	Train	Val	F1-micro
1	1	0.5445	1	1	0.4639
5	5	0.6256	10	10	0.5264
10	10	0.6316	100	100	0.5283
15	15	0.5319	1000	1000	0.5198

C. FEW-SHOT LEARNING WITH DIFFERENT SETTINGS

1) THE NUMBER OF SHOTS

ProtoDAG forms prototypes with a support set to classify queries; the change in number of shots can affect the model performance. Intuitively, when the number of shots are increased, prototypes are expected to be more stable. We observe that the performance improves as the number of shots is increased in Table 6. Interestingly, IEMOCAP performs well even with 1-shot, whereas DailyDialog exhibit poor performance with 1-shot. This is due to the dataset itself, which is summarized in Table 1. The number of conversations of DailyDialog dataset is nearly 100 times many as that of IEMOCAP, and 1-shot is extremely limited considering the size, and the diversity of the conversations.

In the case of IEMOCAP, increasing the number of shots to 5 or 10 can even achieve better performance than some supervised learning models. While better performance can be achieved through increasing the number of shots, in many cases, the number of shots should be limited due to data labeling. When the number of shots increases, it becomes similar to the amount of data used for learning in supervised learning. Therefore, considering ProtoDAG's assumption that there is not much-labelled data, we set the number of IEMOCAP to one shot. We also use a limited number of shots in DailyDialog for the same reason.

2) THE NUMBER OF EPISODES

Table 7 shows how the performance improves from 1 episode to many. When trained over sufficient number of episodes, the model generalizes well. Exploiting only one episode is challenging to generalize to other data that haven't been observed. For IEMOCAP, when the number of episodes are set to 10, the model achieves the best performance. For DailyDialog, when the number of episodes are set to 100, the model achieves the best performance. Although the amount of data in the episode seems small, the model continues to compare episodes of various cases.

On the other hand, when the number of episodes exceeds some limit, the model tend to overfit as shown in Table 7. When we set the the number of episodes to 1000, the maximum used data becomes (5+10)*7*1000 = 105,000, where 5, 10, and 7 are from shots, queries, ways respectively. Since DailyDialog has 10,000 train data, the model could encounters same data in independent support sets and query sets within one epoch. This causes overfitting, which results in performance degradation. From Table 7, it can be seen that the performance degrades on both datasets when the number of episodes is set to 15 and 1000 for IEMOCAP and DailyDialog respectively.

TABLE 8. This table shows performance of supervised learning methods with few data in IEMOCAP dataset.

Model	with full data	with few-data
COSMIC	0.6305	0.3907 (-0.2398)
DialogueRNN	0.6476	0.5145 (-0.1329)
DAG-ERC	0.6803	0.5893 (-0.0910)
ProtoDAG	-	0.6316

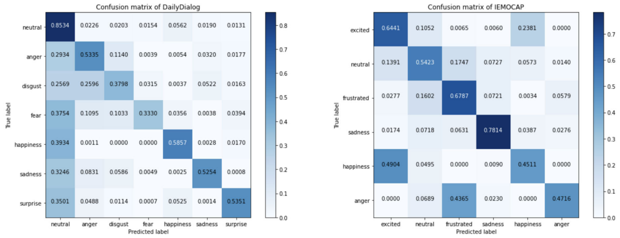


FIGURE 2. The confusion matrix of DailyDialog and IEMOCAP.

D. PERFORMANCE OF SUPERVISED LEARNING APPROACHES WITH FEW-SHOT SETTING

We conduct additional experiments to prove the value of ProtoDAG in Table 8. We feed fewer samples(as less as in the few-shot setting) to supervised models. All supervised model with few-sample shows a significant drop in performance. However, our model outperforms other models with significant margin even though we use same few sample. These results prove that our proposed model is not simply combining the two previous models. We emphasize that for good few-shot performance, few-shot setting is essential.

E. ERROR STUDY

Many previous ERC studies mentioned that ERC's difficulties stem from *emotion shift* and *confusing emotion* [1], [3], [5], [15], [18], [21], [37]. In conversation, utterances often have emotions similar to those of surrounding speech. This phenomenon is *emotion consistency* [38] On the other hand, when each adjacent utterance in a conversation switches from one to another, it is called an *emotion shift*. ProtoDAG's confusion matrix for the DailyDialog dataset also presents difficulties due to emotion shift. In DailyDialog, the most misclassified label of all emotions is neutral. In most cases, the model did not notice the change in emotion well because it was neutral and changed into a different emotion during the conversation.

In the IEMOCAP dataset, the *confusing emotion* phenomenon was more prominent than DailyDialog. It was observed that emotions located in similar areas of emotion, such as excited - happiness and frustrated - anger, were often confused. This phenomenon is also seen in previous studies, but in the case of few-shot learning, this limitation is more evident as the diverse personality of the conversation is ignored. The fundamental reason for the two difficulties in ERC is that emotions are continuous and cannot be explicitly divided.

VI. DISCUSSION

Recently, large language models like ChatGPT [2] have shown significant advancements, excelling in general language understanding and generation tasks. However, these models are primarily pre-trained on massive amounts of generic text data, which can make it challenging for them to accurately extract specific emotions in conversations. Recent work [39] shows that ChatGPT still lags behind supervised models in emotion recognition. Furthermore, these models often require extensive data and have substantial model sizes. Therefore, independent research on tasks such as in emotion recognition is significant, and ongoing efforts are needed.

VII. CONCLUSION

We propose ProtoDAG, which applies DAG in prototypical networks for FSL-ERC. ProtoDAG outperforms previous state-of-the-art FSL method with significant margin. On the DailyDialog dataset, we achieved an improvement in the micro F1-score of 21.02% (from 31.81% to 52.83%); on the IEMOCAP dataset, we achieved an improvement in the weighted F1-score of 25.35% (from 37.81% to 63.16%). ProtoDAG achieved state-of-the-art FSL results. It even improved the FSL performance to a comparable level with the supervised learning methods for ERC.

REFERENCES

[1] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6818–6825.

[2] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," 2020, *arXiv:2010.02795*.

[3] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 1551–1560.

[4] C. Liang, C. Yang, J. Xu, J. Huang, Y. Wang, and Y. Dong, "S+PAGE: A speaker and position-aware graph neural network model for emotion recognition in conversation," 2021, *arXiv:2112.12389*.

[5] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 11595–11603.

[6] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[7] S. M. Zahiri and J. D. Choi, "Emotion detection on TV show transcripts with sequence-based convolutional neural networks," in *Proc. Workshops 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 527–536.

[8] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[10] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. 4th Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, 2017.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 4080–4090.

[12] G. Guibon, M. Labeau, H. Flamein, L. Lefeuvre, and C. Clavel, "Few-shot emotion recognition in conversation with sequential prototypical networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2021, pp. 6858–6870.

[13] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[14] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.

[15] 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 *Proc. Conf. Assoc. Comput. Linguistics North Amer. Chapter. Meeting*, 2018, p. 2122.

[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[17] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," 2019, *arXiv:1904.04446*.

[18] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*.

[19] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7360–7370.

[20] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," 2019, *arXiv:1909.10681*.

[21] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 13789–13797.

[22] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.* Heraklion, Geece: Springer, 2018, pp. 593–607.

[23] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[24] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[25] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, Lille, France, vol. 2, 2015, pp. 1–27.

[27] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.* Copenhagen, Denmark: Springer, 2015, pp. 84–92.

[28] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3637–3645.

[29] T. Gao, X. Han, Z. Liu, and M. Sun, "Hybrid attention-based prototypical networks for noisy few-shot relation classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6407–6414.

[30] Y. Bao, M. Wu, S. Chang, and R. Barzilay, "Few-shot text classification with distributional signatures," in *Proc. ICLR*, 2020.

[2] https://chat.openai.com

[31] S. Sun, Q. Sun, K. Zhou, and T. Lv, "Hierarchical attention prototypical networks for few-shot text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 476–485.

[32] A. Fritzler, V. Logacheva, and M. Kretov, "Few-shot classification in named entity recognition task," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 993–1000.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[34] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.* Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

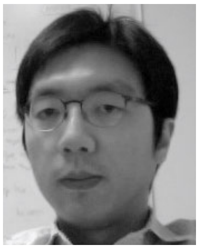
[36] X. Song, L. Huang, H. Xue, and S. Hu, "Supervised prototypical contrastive learning for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 5197–5206.

[37] D. Ghosal, N. Majumder, R. Mihalcea, and S. Poria, "Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 1435–1449.

[38] Y. Wang, J. Zhang, J. Ma, S. Wang, and J. Xiao, "Contextualized emotion recognition in conversation as sequence tagging," in *Proc. 21st Annu. Meeting Special Interest Group Discourse Dialogue*, 2020, pp. 186–195.

[39] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is ChatGPT equipped with emotional dialogue capabilities?" 2023, *arXiv:2304.09582*.

YUJIN KANG received the B.S. degree in Korean language and literature from Chung-Ang University, Seoul, South Korea, in 2022, where she is currently pursuing the M.S. degree in artificial intelligence. Her research interests include natural language processing and emotion recognition in conversation.



YOON-SIK CHO received the B.S. degree in electrical engineering from Seoul National University, South Korea, in 2003, and the Ph.D. degree in electrical engineering from the University of Southern California, USA, in 2014. He was an Academic Mentor of RIPS Program with the Institute for Pure and Applied Mathematics, University of California at Los Angeles, and a Postdoctoral Scholar with the Information Sciences Institute, University of Southern California. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, South Korea. His research interests include large-scale data science, social network analysis, and cloud computing.