An Analysis of Knowledge Distillation Methods for Sentiment Analysis: Limitations and Future Prospects

Qirui Li

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, 215123 Suzhou, China

Qirui.li22@student.xjtlu.edu.cn

Abstract. With the advancement of sentiment analysis, the integration of multi-teacher and multi-modal models has grown increasingly prominent. However, the direct fusion of these diverse models poses significant challenges. This paper presents a comprehensive review of knowledge distillation techniques designed to address these challenges. This paper systematically surveys several key approaches: for multi-teacher scenarios, this paper examines ensemble knowledge distillation (EKD) as a method for effective model compression; for multi-modal applications, this paper explores cross-modal distillation (CMD) and graph-based distillation to effectively handle the intrinsic heterogeneity between data modalities. Besides, this paper analysis identifies critical limitations in current methods, including the static nature of knowledge extraction and constraints of graph distillation approaches. To address these shortcomings, this paper proposes several forward-looking research directions, such as incorporating meta-learning for dynamic weight adjustment, using attention mechanisms for adaptive teacher selection, and replacing conventional Graph Neural Networks (GNNs) with Hypergraph Neural Networks (HNNs) to capture higher-order inter-modal relationships. By critically examining existing methods and a clear outlook on future trends, this paper provides valuable insights and a roadmap for researchers in the field.

Keywords: Knowledge Distillation, Sentiment Analysis, Cross-Modal Distillation, Graph-Based Distillation, Ensemble Knowledge Distillation.

1. Introduction

Sentiment Analysis, a core task within Natural Language Processing (NLP), focuses on the automated identification and classification of emotional tones, attitudes, and opinions expressed in text. The primary objective of this field has evolved from discerning broad polarity (positive, negative, neutral) to more fine-grained analysis of specific affective states, such as joy, anger, and sadness. Methodologically, sentiment analysis has undergone significant evolution: initial approaches in the mid-to-late 20th century were predominantly based on lexicons and rule-based systems; this was followed by a shift towards statistical and machine learning models(e.g., Naïve Bayes and Support Vector Machines), from the 1980s through the early 2000s; the 2010s then saw the rise of deep learning, with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to learning high-level emotional feature representations [1].

The introduction of the "Transformer" architecture marked a significant turning point in 2017. This paradigm, powered by self-supervised learning (SSL), gave rise to revolutionary models such as bidirectional encoder representations from transformers (BERT) and Wav2Vec 2.0 [2]. Their self-attention mechanisms — which weigh the importance of different words within a sequence, enable bidirectional contextual understanding and capture long-range dependencies, both critical for nuanced emotion recognition. To better simulate human perception, the research focus has gradually shifted from single-modal analysis to multimodal emotion recognition (MER), which integrates cues from language, vision, and acoustics. However, the direct fusion of multiple large-scale SSL models poses formidable challenges, including immense computational overhead that is prohibitive for real-time or on-device applications. This has spurred two major lines of inquiry: first, achieving significant model compression without sacrificing performance, through knowledge distillation [3]. Second, effectively manage the inherent heterogeneity between different data modalities effectively through graph distillation [4].

Concurrently, to tackle the fundamental issue of modality heterogeneity, advanced strategies have been developed around the concept of graph distillation: information is decomposed into the shared parts among different modes and the unique parts of each mode through feature decoupling, followed by structured knowledge transfer is carried out. This includes local graph distillation and global graph distillation.

Given the rapid and multifaceted evolution of these techniques, a clear comparative analysis is needed. This paper aims to integrate and synthesize existing research on knowledge distillation for multimodal sentiment analysis. We will systematically compare primary distillation paradigms, evaluating the advantages and disadvantages of single-teacher, multi-teacher, and advanced graph-based distillation methods. By examining core methodologies, key results, and unresolved challenges, this work seeks to consolidate the field's achievements, clarify prevailing trends, and provide a structured reference to guide future research toward developing more efficient, robust, and balanced models.

2. Literature review

2.1. CMD with audio–text fusion

This is a new method that combines audio and text modalities for fine-grained sentiment classification. The aim of this method is to overcome the challenge of limited fine-grained multimodal datasets by leveraging pre-trained BERT (for text) and Wav2vec 2.0 (for audio) as teacher models. The CMD method transfers the feature space knowledge of these large teacher models to a lighter student model (Audio-text transformer), using contrastive learning and dual attention mechanisms (cross-attention and self-attention) [5].

2.2. Decoupled multimodal distilling (DMD)

This is a method for emotion recognition, aiming to address the challenges in multi-modal emotion recognition (MER) where different modalities (such as language, vision, and audio) contribute significantly differently and the inherent multi-modal heterogeneity. By decoupling the features of each modality into modality-independent and modality-exclusive parts, DMD utilizes the graph distillation unit (GD-Unit) to perform flexible and adaptive cross-modal knowledge distillation in these decoupled spaces. DMD enhances the discriminative features of each modality through an autoregressive mechanism and dynamic knowledge transfer [6].

Graph distillation is not a traditional teacher-student model. Instead, it builds a graph to facilitate the knowledge transfer between modality-shared features and modality-specific features. The goal of this distillation is not to compress the model but to narrow the distribution gap between these two types of features, thereby achieving better multi-modal fusion [7]. Essentially, it is a form of knowledge alignment and enhancement across different feature spaces.

2.3. Global distilling framework (GDF)

This method is based on DMD and introduces Modal Cognitive Gravity Loss (MCGL). The aim is to enhance the cognitive correlation of the same modal features across different samples. At the same time, it reduces the cognitive correlation of modal features from different modalities for the same sample [8].

2.3.1 Enhance modality discrimination

Increase the distance between specific features of different modalities to ensure that each modality learns unique information (features), thereby improving the model's discrimination ability.

2.3.2 Reduce intra-modal redundant information

Ensure that the specific features (modality-specific features) of different samples within the same modality can be more concentrated, thereby reducing the repetition or unnecessary information within each modality.

2.4. EKD

This method introduces the implementation of EKD in self-supervised speech teacher models to create smaller, more efficient and comparably performing compressed models. By comparing different aggregation techniques such as hierarchical averaging and hierarchical concatenation, the authors found that these two methods performed poorly in the knowledge distillation scheme. Therefore, they proposed a multi-prediction head approach, allowing the student model to simultaneously predict the outputs of different layers of multiple teacher models [9].

3. Methods

3.1. CMD

3.1.1 Audio-text fusion

The student model incorporates a two-stage attention mechanism within itself, which is used to deeply integrate the information from the two modalities.

Cross-Attention: $f_\theta\left(X_{audio}^q, X_{text}^k, X_{text}^v\right) = C_{audio}$ and $f_\theta\left(X_{text}^q, X_{audio}^k, X_{audio}^v\right) = C_{text}$. This is the first step of integration. The audio converter receives the "keys" and "values" from the text modality, while the text converter receives the "keys" and "values" from the audio modality. This enables the model to learn the alignment and association between audio and text at an early stage.

Self-Attention: $f_\theta\left(C_{audio}^q, C_{audio}^k, C_{audio}^v\right) = C_{audio}^S$ and $f_\theta\left(C_{text}^q, C_{text}^k, C_{text}^v\right) = C_{text}^S$. After cross-attention processing, the feature vectors of each modality are further processed through the self-attention mechanism to further refine their own contextual information.

3.1.2 CMD

By using contrastive learning to train the student model, we make the context vectors generated from the same original data as similar as possible, that is, the student vector Z^s and the teacher vector Z^t are similar.

$$\mathcal{L}_{i,j} = -log\frac{\exp\left(\frac{S_{ij}}{K}\right)}{\sum_{k=1}^{2n}\exp\left(\frac{\mathbb{S}_{i\mathbb{k}}}{K}\right)},\ \mathcal{L} \in \mathbb{r}^{2n\times 2n} \tag{1}$$

In eq. (1), S_{ij} represents the cosine similarity between vectors i and j. The numerator is the similarity score of the positive sample pairs, while the denominator is the similarity score of the negative sample pairs. By minimizing this loss, the model is trained to be able to better distinguish between positive sample pairs and negative sample pairs.

$$\mathcal{L}_{clr}(Z^t, Z^s) = \frac{1}{2n}\sum_{k=1}^{n}[\mathcal{L}_{k,n+k} + \mathcal{L}_{n+k,n}] \tag{2}$$

In eq. (2), $\mathcal{L}_{k,n+k} + \mathcal{L}_{n+k,n}$ this is a symmetric loss. It calculates the loss when the teacher vector Z^t is used as the query and the student vector Z^s is used as the positive sample and also calculates the loss in the reverse direction when the student vector is used as the query and the teacher vector is used as the positive sample. This approach makes the training more stable and the effect better.

$$\mathcal{L}_{total} = \mathcal{L}_{clr}\left(Z_{audio}^t, Z_{audio}^S\right) + \mathcal{L}_{clr}(Z_{text}^t, Z_{text}^S) \tag{3}$$

In eq. (3), the sum of the audio contrast loss and the text contrast loss give $\mathcal{L}_{total}$, and the gradient descent algorithm is used to minimize $\mathcal{L}_{total}$.

3.1.3 Fine-tuning on the downstream task

The sentiment labels were fine-tuned to verify whether the text-audio transformer would help improve the performance of sentiment classification.

The first step is featuring extraction and processing. Original data of audio and text are transformed into context embeddings through the text-audio transformer. Then, by performing average pooling on the time steps, they are converted into fixed-length vectors.

The second step is sentiment classification prediction. The pooled vectors are sent to the fully connected layer separately. Then, the output of the fully connected layer is converted into the prediction probabilities of each sentiment category using the Softmax function.

The third step is calculating Loss and Update Model. The cross-entropy loss between the predicted probabilities and the true sentiment labels is finally calculated. The model is updated using the gradient-decent algorithm.

3.2. DMD

The core of this method lies in the application of two sets of core loss functions:

$$\mathcal{L}total = \mathcal{L}task + \lambda_1 \mathcal{L}dec + \lambda_2 \mathcal{L}dtl \tag{4}$$

In eq. (4), $\mathcal{L}task$ is designed to minimize the error between the emotional predictions during training and the actual values. $\mathcal{L}dec$ is designed for audio-text fusion. $\mathcal{L}dtl$ represents graph distillation loss.

3.2.1 Decoupling loss: $\mathcal{L}dec$

Preparing work: By decoupling the features of multiple modalities, two types of features are extracted: the shared feature X_m^{com}, which is common to different modalities and independent of the modality, and the homogeneous feature X_m^{prt}, which is unique to each modality and heterogeneous. $m \in \{L, V, A\}$ Language (L), Visual (V), Acoustic (A).

The first set of loss: By decomposing the features of each modality into two parts: modality-independent and modality-exclusive and using "autoregressive" to perform a "split-merge-reconstruct-repredict" cyclic process, the aim is to ensure that the model can effectively distinguish the shared information and exclusive information between different modalities, thereby achieving a more thorough feature decoupling. The components of its loss function are:

$$\mathcal{L}dec = \mathcal{L}rec + \mathcal{L}cyc + \gamma(\mathcal{L}mar + \mathcal{L}ort) \tag{5}$$

In eq. (5), $\mathcal{L}dec$ consists of four parts. $\mathcal{L}rec$ represent reconstructing loss. $\mathcal{L}rec$ represents cycle loss. $\mathcal{L}mar$ represents marginal loss. $\mathcal{L}ort$ represents soft orthogonal loss.

Reconstructing loss: $\mathcal{L}rec = \left\| \tilde{X}_m - D_m([X_m^{com}, X_m^{prt}]) \right\|_F^2$ this loss function is used to measure the difference between the features re-combined from the two decoupled parts (mode-independent and mode-specific) and the original features.

Cycle loss: $\mathcal{L}cyc = \left\| X_m^{prt} - \varepsilon_m^{prt}(D_m([X_m^{com}, X_m^{prt}])) \right\|_F^2$. this loss function is used to verify the retention of heterogeneity.

Marginal loss: $\mathcal{L}mar = \frac{1}{|S|} \sum_{(i,j,k) \in S} \max(0, \alpha - \cos(\mathcal{X}_{m[i]}^{com}, \mathcal{X}_{m[j]}^{com}) + \cos(\mathcal{X}_{m[i]}^{com}, \mathcal{X}_{m[k]}^{com}))$ this loss function indicates that homogeneous features from the same emotion, but different modalities should be more similar than those from the same modality but different emotions.

Soft Orthogonal Loss: $\mathcal{L}ort = \sum_{m \in \{L,V,A\}} \cos(\mathcal{X}_m^{com}, \mathcal{X}_m^{prt})$ this loss function aims to make the common information of the modes and the independent information of the modes as independent as possible, with a cosine similarity of 0 meaning orthogonality.

3.2.2 Graph distillation loss: $\mathcal{L}dtl$

The second group of losses: This is the core loss function of the GD-Unit. It is composed of the element-wise product of the distillation strength matrix W and a distillation difference matrix E that represents the differences between different modes.

$$\mathcal{L}dtl = \|W \odot E\|_1 \tag{6}$$

In eq. (6), $E = \varepsilon_{i\to j}$. E represents the distillation matrix, which records the distillation differences between all modalities. Specifically, it is the difference in logits between v_i and v_j, where v_i's logits is derived from (x_i, θ_1).

$$W = \omega_{i\to j} = g([f(x_i, \theta_1), x_i], [f(x_j, \theta_1), x_j]\theta_2) \tag{7}$$

In eq. (7), f is a fully connected layer (FC layer) with learning parameters θ_1. It is used to regress logits from the features. g is another FC layer with learning parameters θ_2. After receiving two concatenated features, it outputs the distillation strength $\omega_{i\to j}$.

The dynamic learning lies in that the distillation strength $\omega_{i\to j}$ from modality i to modality j is not a preset parameter. Instead, it is dynamically calculated through a fully connected layer g based on the current input modality features (x_i, x_j) and their logits.

Homogeneous graph knowledge distillation (Homo GD): Feature $\mathcal{X}_m^{com}$ is put into x_i in eq. (7) and calculate $\mathcal{L}_{dtl}^{homo}$ using eq. (6).

Heterogeneous graph knowledge distillation (Hetero GD): Its function is to receive features from different modalities and perform cross-modal information fusion, thereby establishing adaptability between modalities and bridging the distribution gap.

Taking the language modality $\mathcal{X}_L^{prt}$ as the source, and the visual modality $\mathcal{X}_V^{prt}$ as the target. Based on the concept of Transformer, the cross-modal attention can be defined as: $Q_V = \mathcal{X}_V^{prt} P_q$ (Query, Q); $K_L = \mathcal{X}_L^{prt} P_K$ (Key, K); $V_L = \mathcal{X}_L^{prt} P_V$ (Value, V).

Then, through cross-modal attention calculation: $\mathcal{Z}_{L\to V}^{prt} = \text{softmax}(\frac{Q_V K_L^T}{\sqrt{d}})V_L$, generate heterogeneous features that enhance from the language module to the visual module. Next, $\mathcal{Z}_{L\to V}^{prt}$ is put into x_i in eq. (7), and $\mathcal{L}_{dtl}^{hetero}$ is calculated by eq. (6).

In final, $\mathcal{L}dtl = \mathcal{L}_{dtl}^{homo} + \mathcal{L}_{dtl}^{hetero}$.

3.3. GDF

$$\mathcal{L}total = \mathcal{L}task + \lambda_1 \mathcal{L}dec + \lambda_2 \mathcal{L}dtl + \lambda_3 \mathcal{L}mcg \tag{8}$$

In eq. (8), the concepts of $\mathcal{L}task, \mathcal{L}dec, \mathcal{L}dtl$ are same in eq. (4). The aim of $\mathcal{L}mcg$ is to enhance the similarity of the same modality and improve the discrimination ability for different modalities.

3.3.1 MCGL

$$\mathcal{L}mcg = \alpha F_1 + e^{-F_2} \tag{9}$$

In eq. (9), F_1 is designed for distinguishing the ability of different modality. e^{-F_2} is designed for improving similarity between features of same modality.

$$F_1 = \sum_1^n (F_{s(h,k)}^{Inter}(X_i) + F_{s(h,z)}^{Inter}(X_i) + F_{s(k,z)}^{Inter}(X_i)) \tag{10}$$

$$F_1 F_{s(h,k)}^{Inter}\left(X_i^{s(h)}, X_i^{s(k)}\right) = G\frac{I(X_i^{s(h)}, X_i^{s(k)})}{\left\|X_i^{s(h)} - X_i^{s(k)}\right\|_2^2}, s \in \{L, A, V\} \tag{11}$$

Eq. (10) means the sum of all inter modal cognitive.

Eq. (11) represents inter modal cognitive. $X_i^{s(h)}, X_i^{s(k)}$ represent modality-specific features of different modalities $s(h)$ and $s(k)$. By minimizing F_1 and thereby weakening the gravitational

pull between different modalities, each modality learns unique information (features) that are more discriminative.

$$F_2 = \sum_1^n \sum_{s\in(L,A,V)} F^{Intra}(X_i, X_j) \tag{12}$$

$$F_{s(h)}^{Intra}\left(X_i^{s(h)}, X_j^{s(h)}\right) = G\frac{I(X_i^{s(h)}, X_j^{s(h)})}{\left\|X_i^{s(h)} - X_j^{s(k)}\right\|_2^2}, s \in \{L, A, V\} \tag{13}$$

Eq. (12) means the sum of all modal internal gravities. By maximizing F_2, the gravitational force between the same-modal features is enhanced, making the features within the same mode more similar and clustered, and reducing redundant information.

Eq. (13) represents intra modal cognitive of sample i and sample j in one modality $s(h)$.

3.4. EKD

3.4.1 Knowledge distillation

The essence of knowledge distillation is for the student mode(S) to learn the output of the teacher model (T).

$$\mathcal{L}_i\left(h_i^S, h_i^T\right) = \frac{1}{D_T}\left\|h_i^S - h_i^T\right\| - log_\sigma(cossim\left(h_i^S, h_i^T\right)) \tag{14}$$

In eq. (14), D_T represents the feature dimension output by the teacher model. $\left\|h_i^S - h_i^T\right\|$ transforms the hidden layer output of student model D_S into the same dimension D_T as that of the teacher model.

Then, using L1 Loss, it calculates the difference between the student's prediction result (h_i^S) and the teacher's target (h_i^T).

For $-log_\sigma(cossim\left(h_i^S, h_i^T\right))$, it utilizes the cosine similarity loss function to ensure the consistency of the directions of the teacher and student feature vectors.

$$\mathcal{L} = \frac{1}{3}\sum_{i\in\{4,8,12\}} \mathcal{L}_i\left(h_i^S, h_i^T\right) \tag{15}$$

Eq. (15) selects multiple hidden layers, enabling the student model to learn more comprehensive and hierarchical information, thereby achieving better generalization.

3.4.2 Ensemble of teacher model

The key is to set up multiple prediction heads. The original model of each teacher model is retained, and multiple sets of prediction heads are used after the last transformer encoder layer of the student model.

The prediction for each teacher is derived by transforming the output of the last hidden layer of the student model through a single prediction head.

$$\mathcal{L}_{multi-pred} = \frac{1}{3}\cdot\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in\{4,8,12\}} \mathcal{L}_i\left(h_i^{S_m}, h_i^{T_m}\right) \tag{16}$$

In eq. (16), $h_i^{S_m}$ represents the prediction of the head for the m-th group, while $h_i^{T_m}$ represents the prediction of the hidden layer output of the m-th teacher model.

4. Limitations

4.1. Limitations of CMD

While the student model achieves outstanding performance in multimodal sentiment classification tasks, its design has fundamental limitations that restrict its application potential in more complex scenarios. First, the CMD framework assumes a static and fixed knowledge hierarchy. Despite the student model's modality fusion through cross-attention within itself, the way audio and text mutual reference mechanism between audio and text modules ignores the dynamics of emotional expression

and certain context dependencies. Second, the CMD framework suffers from information bottlenecks. Although the student model achieves modality fusion through cross-attention within itself, the teachers' knowledge remains isolated. BERT only provides text knowledge, and Wav2vec 2.0 only provides audio knowledge [2]. This framework lacks a more powerful "super teacher" that has already integrated multimodal information. Therefore, the student model cannot directly learn how an advanced model represents the interaction between modalities and how to balance the weights of different modalities in the interaction. It can only explore this interaction within its own limited parameter space under the guidance of two isolated knowledge sources, which undoubtedly increases the difficulty and uncertainty of learning.

Upon a thorough analysis of the limitations of CMD, it can be found that the text modality often takes the leading role in many scenarios as it directly carries semantic information. However, it must also be acknowledged that for fine-grained emotion classification, the audio modality is indispensable. As literature notes: "many emotions are easier to recognize through language, while others are easier to recognize through vision." In some cases, the audio modality may contain more crucial emotional information than the text. A typical example is irony: the text content might be "Well done!", but the sarcastic tone completely subverts its literal meaning. In such cases, the audio is the dominant modality. The static framework of CMD is unable to adapt to the dynamic changes in the importance of modalities based on the context.

Therefore, a static distillation framework with a predefined direction of knowledge flow has fundamental flaws. It cannot adapt to situations where a "weaker" modality holds the most critical information at a specific moment. This insight necessitates a shift toward a dynamic distillation framework, in which the influence of each modality teacher can be weighted and adjusted in real time. This is precisely the core of the future research direction to be explored in the third part.

4.2. Limitations of graph neural networks

4.2.1 Structural constraints and pairwise restrictions

The structure of the GD-unit is defined by a graph, which contains four central vertices: one representing the aggregated modality-shared features (V_c), and three representing the modality-specific features of language, vision, and audio (V_c, V_V, V_A). The edges connecting these vertices define the paths and intensities of knowledge transfer (refinement).

However, this architecture is based on the principles of traditional GNNs, with the refinement paths mainly limited to sequential or parallel binary interactions (e.g., $V_C \leftrightarrow V_V$, $V_c \leftrightarrow V_A$, $V_V \leftrightarrow V_A$). This design limits the ability of the GD-unit to capture complex cross-modal dependencies.

4.2.2 Deficiencies in high-order emotional interaction modeling

The key flaw of this pairwise structure becomes apparent in high-order interaction scenarios commonly seen in conversational emotion recognition in context (CERC). Emotional states, especially in natural conversations, are typically conveyed through synchronous, nonlinear interactions of three or more modalities. For example, expressions such as irony or emotional conflict require the model to be able to recognize the inconsistency where the language (L) semantics may be positive, but the acoustic (A) pitch and visual (V) cues are negative. This demands explicit modelling of collective, represented as L∩V∩A dependencies, which exceeds the scope of standard pairwise GNNs.

The fixed 4-node topology of the GD-unit, although highly efficient in feature alignment, lacks the necessary structural complexity to define and learn the required dependencies. This structural limitation may lead to poor fusion effects when dealing with highly contextual, complex, and high-order emotional signals [10].

4.3. Limitations of EKD

The core of the EKD framework is to enable a student model to learn simultaneously from a "committee" composed of multiple teacher models. These teacher models can be of different

architectures, trained on different datasets, or models with different noise robustness on the same dataset.

The authors compared various aggregation strategies and proposed a method superior to layerwise-averaging/concatenation: Multiple Prediction Heads. This method equips each teacher model with a set of independent prediction heads. The total distillation loss is the weighted sum of all prediction head losses. Experiments confirm that this method performs best because it avoids the destructive interference between different teacher signals and allows the student model to independently learn knowledge from each expert.

Although the multiple prediction head method is effective, it is essentially a parallel, non-interactive learning paradigm. The student model learns to imitate each teacher independently, but the entire framework lacks an intelligent mechanism to integrate or prioritize these potentially conflicting or overlapping knowledge. When different teachers give different "opinions" for the same input, the student model can only passively accept these mixed signals through simple loss summation. As noted in relevant research, this requires a more advanced method to coordinate the diverse knowledge from different teachers. The student model does not learn "under what circumstances it should trust which teacher more", which limits its learning depth and efficiency.

5. Suggestions

5.1. Beyond fixed weights : meta-learning for teacher prioritization in CMD

To address the static knowledge flow limitation of the CMD framework, future distillation frameworks must be capable of dynamically adjusting the importance of knowledge sources. A promising direction is the integration of meta-learning. Specifically, a small meta-network can be constructed whose input is the current data sample (or its shallow representation) and whose output is a set of weights. In the context of multi-modal distillation, this set of weights will dynamically allocate the distillation loss to each modality.

For instance, when the input is speech with a distinct sarcastic tone, the meta-network should learn to assign a higher weight to the audio teacher, such as WavLM, while reducing the weight of the text teacher, such as BERT. Conversely, when the input is a text with explicit emotional expression and complex context, the weight of the text teacher should be increased. This transforms the distillation process from a fixed and unchanging procedure to an adaptive, data-driven system. The student model no longer passively absorbs all knowledge but dynamically selects dynamically choose which "expert" (teacher model) to follow based on the nature of the current problem.

5.2. Hypergraph-enhanced global graph refinement unit (HG-GGD)

To address the limitations of traditional GNNs in the GD-unit pairwise relationship limitation, HNNs can create a HG-GGD.

The reason why hypergraph neural networks are helpful in enhancing the modelling of emotional propagation is that they overcome the inherent limitations of traditional GNNs, namely the limitation of pairwise relationships. HNNs can construct hyperedges, thereby simultaneously connecting multiple modalities or utterance nodes. This ability enables the model to capture complex connection patterns where three or more dialogue elements or modality features simultaneously influence emotions. By modelling these complex relationships, HNNs can enhance the model’s ability to capture emotional propagation, which is crucial for accurately identifying the emotional state of each discourse in a conversation.

The graph propagation step of the standard GD-unit can be replaced or supplemented by a HNNs module. This HNNs can not only aggregate information flow through traditional pairwise edges but also through newly defined hyperedges. Then, the distillation loss will be calculated based on the output of this Hypergraph-enhanced GNNs, ensuring that knowledge transfer accounts for the complex, simultaneous interactions between all modalities. This process enables the shared feature

representation to be learned not only to be a summation of pairwise alignments, but also to be a true distillation of the collective and high-level emotional context.

By achieving this dynamic modelling of high-level dependency relationships, HG-GGD can significantly enhance the modelling ability of emotional propagation in complex situations, such as those involving multiple speakers or conflict signals, thereby improving overall performance.

5.3. EKD

EKD can shift from simple "imitation" to a more advanced "selective integration". It can use cosine similarity to dynamically select the teacher model whose output is most consistent with the current student model's output.

The logical chain of this evolutionary process is as follows: (1) The initial EKD paradigm treats all teachers as equally valid sources of truth, and the student needs to learn all knowledge indiscriminately; (2) However, reality is that different teachers have different specialties. For example, one teacher is good at processing clean speech, while another is more robust in noisy environments. For a specific input (such as a noisy speech segment), the opinion of one teacher is obviously more reliable than that of the other; (3) Therefore, the student should not learn equally from all teachers. It should learn to judge which teacher to trust based on the current context.

This means that the core of multi-teacher distillation in the future lies in directly incorporating a "gate" or "attention" mechanism into the distillation loss function. The student model not only needs to predict the output of the teacher but also should simultaneously predict a set of weights, which represent the stronger relevance and reliability for the current input. This will transform knowledge distillation from a passive imitation process into an active, selective knowledge integration process.

6. Conclusion

This paper systematically reviews three key methods of knowledge distillation in the field of sentiment analysis: CMD, EKD of self-supervised speech models, and multimodal knowledge distillation via graph-based frameworks (GDF/DMD). Through the review and analysis of existing literature, this paper finds that while each method has unique advantages, they all have significant limitations in terms of the static nature of the distilled models and constraints in graph-based distillation. Specifically, CMD fails to dynamically adjust the teacher weights based on emotional expression and contextual nuances. GDF/DMD is limited by the pairwise relationships between nodes in GNNs, while EKD is hindered by its parallel, non-interactive learning paradigm. These limitations not only restrict the application effect and scope of existing methods but also highlight directions for future breakthroughs.

To address these challenges, this paper proposes future research directions: Introducing meta-learning to dynamically adjust the weights of teacher models, replacing GNNs with HNNs to construct hyperedges and simultaneously connect multiple modalities, integrating attention mechanisms to dynamically select teachers whose outputs align most closely with the current student model. This paper expects that through these explorations, the bottlenecks of current methods can be effectively overcome, bringing new breakthroughs to the field of knowledge distillation in sentiment analysis.

In future work, this paper aim to translate these theoretical improvement suggestions into specific code implementations and experimental validations. This paper will conduct empirical research to test the feasibility and practical effects, particularly in enhancing of the proposed methods in improving the accuracy of complex emotion recognition and operational efficiency.

References

[1] M. Kamyab, G. Liu, Michael. A, Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis. Appl. Sci. 11, 11255 (2021).

[2] D. Kounadis, O. Schrüfer, A. Derington, H. Wierstorf, F. Eyben, F. Burkhardt, B. Schuller, wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems. 33, 12449-12460, (2020).

[3] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network. NIPS workshop. (2015).

[4] Y. Tsai, S. Bai, P. Liang, J. Kolter, L. Morency, R. Salakhutdinov, Multimodal Transformer for Unaligned Multimodal Language Sequences. In proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July (2019), 6558-6569.

[5] D. Kim, P. Kang, Cross-modal distillation with audio–text fusion for fine-grained emotion classification using BERT and Wav2vec 2.0, Neurocomputing. 506, 168-183 (2022).

[6] Y. Li, Y. Wang, Z. Cui, Decoupled Multimodal Distilling for Emotion Recognition. In proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, September 13 (2025), 6631-6640.

[7] Z. Luo, J. Hsieh, L. Jiang, J. Niebles, F. Li, Graph Distillation for Action Detection with Privileged Modalities. In proceedings of the European Conference on Computer Vision, Sep (2018), 166-183.

[8] H. Li, H. Zhong, C. Xu, X. Liu, G. Wen, L. Liu, Global distilling framework with cognitive gravitation for multimodal emotion recognition. Neurocomputing, 622, 129306 (2025).

[9] K. Huang, T. Feng, Y. Fu, T. Hsu, P. Yen, W. Tseng, K. Chang, H. Lee, Ensemble Knowledge Distillation of Self-Supervised Speech Models. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 4 (2023), 1-5.

[10] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph Neural Networks, In proceedings of the AAAI Conference on Artificial Intelligence. Jul 17 (2019), 3558-3565.