論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Empowering Emotion Recognition with Flexible Modality Information
著者(和文)	東遠 李
Author(English)	Dongyuan Li
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12860号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:奥村 学,中山 実,鈴木 賢治,篠﨑 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12860号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信系　　系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工　）
学生氏名： Student's Name	LI DONGYUAN	審査員主査： Chief Examiner	奥村　学　教授

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Teaching machines to understand human emotion is one of the most elusive and long-standing challenges in Artificial Intelligence. This thesis tackles two core issues of emotion recognition: (1) how to effectively and efficiently apply unimodal emotion recognition tools in real-world scenarios; (2) how to build a general multimodal emotion recognition model with high performance. Specifically, we focus on unimodal and multimodal emotion recognition: a class of emotion recognition models built on top of deep neural networks. Compared to traditional sparse, hand-designed feature-based machine learning methods or statistic models, these end-to-end neural models have proven to be more effective in learning and extracting rich sentiment and semantic information and improved performance on all modern emotion recognition benchmarks by a large margin. This thesis consists of two parts. In the first part, we aim to cover the essence of unimodal emotion recognition and present our efforts at building effective and efficient unimodal emotion recognition models. Specifically, existing unimodal emotion recognition methods often overlook the information gap between the pre-trained models and the downstream emotion recognition task, resulting in sub-optimal performance. Moreover, current methods require much time for fine-tuning on each specific unimodal dataset, which limits their effectiveness in real-world scenarios with large-scale noisy data. To address these issues, we take speech emotion recognition as an example and propose an active learning (AL)-based fine-tuning framework for speech emotion recognition, called After that leverages task adaptation pre-training (TAPT) and AL methods to enhance performance and efficiency. Specifically, we first use TAPT to minimize the information gap between the pre-trained speech recognition task and the downstream speech emotion recognition task. Then, AL methods are employed to iteratively select a subset of the most informative and diverse samples for fine-tuning, thereby reducing time consumption. Experiments demonstrate that our method After, using only 20% samples, improves precision by 8.45% and reduces time consumption by 79%. The additional extension of After and ablation studies further confirms its effectiveness and applicability to various real-world scenarios. We also summarize limitations and discuss future directions in this field. In the second part of this thesis, we aim to cover the essence of multimodal emotion recognition and present our efforts at building effective and robust multimodal emotion recognition models. Specifically, graph-based multimodal emotion recognition models have achieved state-of-the-art performance on multiple benchmarks. However, current graph-based methods fail to simultaneously depict global contextual features and local diverse unimodal features in a dialogue. Furthermore, with the number of graph layers increasing, they easily fall into over-smoothing. In this paper, we propose a method for joint modality fusion and graph contrastive learning for multimodal emotion recognition (Joyful), where multimodal fusion, contrastive learning, and emotion recognition are jointly optimized. Specifically, we first design a new multimodal fusion mechanism that can provide deep interaction and fusion between the global contextual and unimodal specific features. Then, we introduce a graph contrastive learning framework with inter-view and intra-view contrastive losses to learn more distinguishable representations for samples with different sentiments. Extensive experiments on three benchmark datasets indicate that Joyful achieves state-of-the-art performance compared to all baselines. We also summarize recent advances and discuss future directions and open questions in this field.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).