Remarkably, across all 20 intent categories, our approach attains top-2 scores in 13 categories, comprising 7 highest scores and 6 second highest scores, which indicates that TCL-MAP achieves better performance than the majority of baselines across various classes. Specifically, in categories like"Complain", "Agree" and "Leave", TCL-MAP consistently outperforms the best baseline by over 1%. Significantly, the "Leave" category exhibits the most substantial improvement of 7.63%. The significant gains can be attributed to TCL-MAP's utilization of modality-aware prompts for better text representation, which in turn enhances video and audio learning through token-level contrastive learning. Nevertheless, in the "Taunt" and "Joke" categories, TCL-MAP seems to provide less assistance in recognizing the intent, which could be caused by a combination of factors, including the limited availability of data within these categories and the intricate nature of the intents themselves. On the other hand, we evaluate the efficacy of TCL-MAP in comparison to the human performance. From the results, we can observe that humans achieve the best performance in the majority of intent categories, which confirms the strong ability of humans to process multimodal information and infer intents through them. However, TCL-MAP surpasses human performance in the "Apologize," "Thank," and "Agree" classes, showcasing the stability of our method when handling challenging samples where humans may make mistakes. In addition, TCL-MAP has approached human performance in intent categories (e.g. "complain", "praise" and "care") which involve distinct emotional aspects and also achieved comparable performance to humans in intent categories (e.g. "Inform", "leave" and "prevent") which require an understanding of actions. These findings further validate the capability of TCL-MAP to effectively extract fearures related to human intents from raw multimodal data, such as expressions, tone of speech and movements.

Comparison between Handcraft Prompt and Modality-Aware Prompt

To further analyze the superiority of our modality-aware prompt, we conduct experiments with handcrafted prompt and modality-aware prompt respectively. Concretly, we select the MIntRec dataset for our experiments, driven by the fact that certain labels (e.g. "Others") in the MELD-DA dataset do not strictly represent intent categories. To make comparison, we design two handcraft prompts aimed at expressing ideas or intents, "I want to" and "I intend to", which maintain the same positions and lengths with the modality-aware prompt. Besides, we conduct an additional set of experiments using [MASK] as the prompt to demonstrate the effectiveness. As shown in Figure 3, we observe a substantial performance advantage in the model that employs the modality-aware prompt in comparison to models using handcrafted prompts, thanks to better integration of non-textual modalities enhancing textual intent semantics extraction. Conversely, the [MASK] prompt shows a notable performance decline compared to handcrafted prompts, highlighting the risk of inappropriate prompts misleading intent understanding. Our modality-aware prompt incorporates the instance-conditional prompt concept of CoCoOp (**?**), thereby mitigating this drawback.

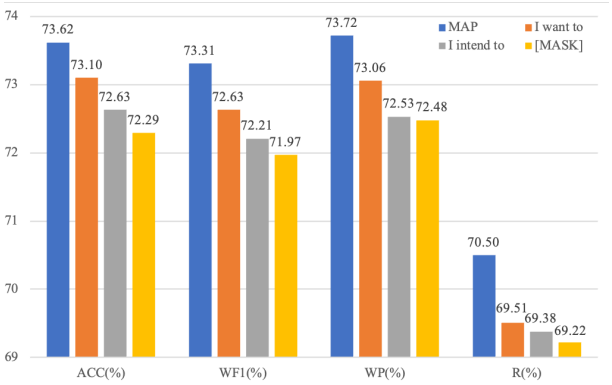

Figure 3: The comparison between Handcraft Prompt and Modality-Aware Prompt

Conclusion

In this paper, we propose a novel Token-Level Contrastive Learning with Modality-Aware Prompting (TCL-MAP) method for multimodal intent recognition. By strengthening the correlations among modalities, our method generate the modality-aware prompt to construct an optimal multimodal semantic space for enhancing the refinement of the text modality. In return, the attained textual representation, enriched with semantics from the ground truth label token, guides the learning process of nonverbal modalities through the token-level contrastive learning. Extensive experiments on two benchmark datasets demonstrate that our approach outperforms state-of-the-art methods and carries significant implications for multimodal prompt learning.

Acknowledgements

This work is founded by the National Natural Science Foundation of China (Grant No. 62173195), National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (Grant No.20232ABC03402), High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City (Grant No. Hongke Zi (2022) 321-16), and Natural Science Foundation of Hebei Province, China (Grant No. F2022208006). Magnam unde repudiandae corrupti, error consequatur provident hic adipisci exercitationem et libero esse, fugit architecto iure nisi minima aut dicta reiciendis, beatae eveniet quidem error nihil unde neque ipsam quasi ipsa repellendus officia?Molestias animi nobis velit beatae est itaque culpa ipsa saepe nemo voluptatum, nisi numquam totam minus reprehenderit explicabo?Veniam adipisci atque itaque minima, pariatur nihil minima saepe unde suscipit animi nam rem aperiam, nam explicabo totam omnis, eaque ab molestias ipsum iure facere veritatis accusantium nam, fuga quis laboriosam excepturi suscipit assumenda provident et consequuntur illo at deleniti.Excepturi ullam quo atque earum veniam alias omnis ipsa, veniam quaerat in quibusdam dolor suscipit modi quae labore accusantium, placeat vitae voluptates laboriosam quae

cumque officiis?