Review of Few-short Multimodal Spoken Language Understanding

Hantian Long
SYSU
htlong21@mail.sysu.edu.cn

Haihen Liu
SYSU
haihlliu21@mail.sysu.edu.cn

Abstract

This work provides a brief overview of the topic of the essay, which is Few-short Multimodal Spoken Language Understanding (FS-MLU). FS-MLU is a field of research that aims to develop conversational systems capable of comprehending and responding to human communication across multiple modalities such as speech, gestures, and facial expressions. We highlights that conventional conversational systems fail to capture the complexities of human communication, and hence FS-MLU is an essential component of everyday life. This work aims to provide an in-depth analysis of the concepts, approaches, and challenges associated with FS-MLU research. It also highlights that recent research has focused on developing FS-MLU systems that leverage machine learning algorithms to comprehend and respond to human communication across multiple modalities. Additionally, we briefly touches on the importance of the essay, which is to provide insights into the field of FS-MLU, its potential impact in revolutionizing human-computer communication, and the challenges that researchers face in its development. In summary, this work provides a concise overview of the essay's main points, the importance of the topic, and what the reader can expect from reading the essay.

1 Introduction

With the rise of the digital world, human-computer communication has become an essential part of everyday life. However, conventional conversational systems fail to capture the complexities of human communication, which involves multiple modes of expression such as speech, gestures, and facial expressions. To address this limitation, recent research has focused on developing few-short multimodal spoken language understanding (FS-MLU) systems, which leverage machine learning algorithms to comprehend and respond to human communication across multiple modalities. This essay aims to provide an in-depth analysis of the concepts, approaches, and challenges associated with FS-MLU research.

FS-SLU is a new domain of Natural Language Processing (NLP) which aims to develop intelligent systems that can comprehend and process human language during interactive conversations. The primary objective of FS-SLU is to develop models that can efficiently generalize to new classes with very few examples (typically less than five), and a small number of classes. Unlike traditional models for spoken language understanding that require a large number of training examples from many classes, FS-SLU models can be trained with very little data and by leveraging pre-trained models.

The field of FS-SLU is gaining momentum due to its ability to generalize to new scenarios, languages, and dialogues. Many researchers are actively working on creating prototype models using various approaches, including Meta-learning, Transfer learning, Data augmentation, and Neural architecture search. One of the primary challenges of FS-SLU is the scarcity of data. Typically, a few-shot learning model will require hundreds if not thousands of samples to be accurate. To address this challenge, researchers have experimented with different techniques such as data augmentation, unsupervised pre-training, and meta-learning.

Another challenge of FS-SLU is the diversity of the conversations and the need for contextual understanding. Unlike traditional NLP models that deal with fixed texts and well-defined applications, FS-SLU models must process and understand the context of the conversation, and the speaker's intent. FS-SLU can have many applications, including text-to-voice (TTS) and speech-to-text (STT) systems, chatbots, Virtual Assistants, and many other interactive systems. The ability to decrease the number of training samples in SLU systems has the potential to open up new domains of human-computer interaction, including tasks that have been impossible to solve before.

2 Related work

2.1 Spoken Language Understanding

Spoken Language Understanding (SLU) is a crucial aspect of human communication. It is the process of understanding the meaning of spoken language by analyzing the words, phrases, and context in which they are used. SLU is used in many areas, including speech recognition, machine learning, and natural language processing.

SLU is particularly important in the field of voice assistants, chatbots, and other forms of automated communication. These technologies rely on SLU to understand user requests, queries, and commands, and to provide accurate and relevant responses. Without SLU, voice assistants would struggle to comprehend the intent behind a user's words and could not provide useful information or complete tasks effectively.

In addition to its significance in the field of automation, SLU is a vital tool for human communication. For example, in healthcare, SLU can be used to analyze a patient's speech patterns, tone, and word choice to identify potential indicators of underlying health conditions. In call centers, SLU can help agents understand customers' inquiries and concerns more accurately and efficiently, leading to better customer service experiences.

Furthermore, SLU can be used to overcome language barriers by enabling accurate translation of spoken language. This is especially important in multicultural and multilingual societies where communication is essential for social and economic integration. SLU can facilitate communication between people who speak different languages by accurately translating spoken language in real-time.

SLU also has significant implications for fields such as education and language learning. By analyzing spoken language, SLU can provide insights into a learner's progress and performance, helping educators tailor their teaching methods and materials to meet individual needs. Additionally, SLU can be used to create personalized language learning experiences that adapt to the student's individual learning style and pace.

Overall, Spoken Language Understanding plays a critical role in modern communication, automation, healthcare, education, and language learning. Its ability to analyze spoken language and provide accurate and relevant responses is essential for improving communication, breaking down language barriers and creating more inclusive and accessible communities. With the continued development of technology, SLU is likely to become even more advanced and essential in the future, providing new opportunities for improved communication and understanding between people and machines.

2.2 Multimodal spoken language understanding

Multimodal spoken language understanding (SLU) is a crucial aspect of natural language processing (NLP) that has gained significant attention in recent years. The ability to extract meaning from speech, text, and other non-verbal cues is the cornerstone of intelligent virtual assistants, chatbots, and other conversational agents. However, the traditional approach to SLU, which involves training models on large datasets of transcribed speech and textual annotations, has several limitations. Firstly, it requires a significant amount of labeled data, which is expensive and time-consuming to acquire. Moreover, it often results in models that are brittle and cannot generalize well to new domains or languages, making it challenging to scale up these systems.

In response to these challenges, researchers have proposed a new paradigm in SLU called few-shot or few-short multimodal learning. This approach involves training models on a small number of examples (i.e., few-shot) or a limited amount of data (i.e., few-short) rather than large datasets. It involves combining multiple modalities, such as speech, text, and vision, to make sense of the user's input. Multimodal SLU leverages the complementary information available from different modalities, improving the overall accuracy and robustness of the system.

One of the earliest works in few-shot multimodal SLU was proposed by Raza and Stiefelhagen (2019). They proposed a model that learns to use a small number of speech and text samples to perform multilingual intent classification. The model's design involved text and speech embeddings and an attention mechanism that allows the model to focus on the most informative parts of the input. The results showed that their model outperformed the baselines on a cross-lingual intent classification task.

Another example is the work of Kang et al. (2020), who proposed a few-shot learning-based approach to emotion recognition in speech. The model learns to classify emotions from speech representations using a small number of labeled examples. Their approach is based on a pre-trained transformer model that can encode speech features as well as textual context. The results demonstrated the efficacy of the proposed model in capturing emotional information from speech, outperforming existing state-of-the-art models.

3 Approaches

3.1 Transfer Learning

Transfer Learning is a common approach in machine learning for developing models that are capable of learning from different but related domains. In the context of FS-MLU, this approach involves pre-training models for different modalities, such as speech recognition and vision models for gesture and facial recognition. Then, the pre-trained models are fine-tuned together to enable FS-MLU. This approach is beneficial as it allows researchers to leverage existing models and datasets and reduce the sample complexity required for developing new models. Moreover, it enables researchers to build models that can handle multiple modes of inputs simultaneously without requiring an extensive training dataset for each modality. However, one of the biggest challenges of Transfer Learning is selecting the best pre-trained models and fine-tuning strategies. This is because the pre-trained models may not work well together, and finding the right combination of pre-trained models and training a model on multiple domains can often be challenging.

3.2 Data Augmentation

Data Augmentation is another approach to enhance the performance of FS-MLU models. In this approach, the models are trained on slightly different but similar datasets, such as data collected from another environment, to improve the models' robustness. Data Augmentation aims to complement the limited training data of FS-MLU models, which makes it difficult for them to achieve high performance without overfitting. By training the models on data that have slight variations from the original data, the model becomes more resilient to changes in the input data, leading to better performance. However, one of the challenges of Data Augmentation is generating realistic data sets that capture the variations of real-world scenarios. Moreover, during the process of augmentation, it is essential to avoid any spurious patterns or artifacts that may cause the model to overfit or generalize poorly.

Overall, Transfer Learning and Data Augmentation, are essential in enhancing the development of Few-short Multimodal Spoken Language Understanding systems. While these approaches are still in their beginning stages, FS-MLU has the potential to revolutionize human-computer communication by enabling conversational agents to interpret and respond to multimodal input. With continual research and development, we can expect more practical implementations of FS-MLU models in conversational systems. This can pave the way for more intuitive human-computer interactions in the future.

4 Challenges

Like every developing field, there are several challenges that FS-MLU researchers face. One of the main challenges is a lack of standardized data sets and evaluation metrics, which makes it difficult to compare the results from different studies. Additionally, there is a challenge around implementing real-time multimodal inputs in practical conversational systems. These systems require quick processing, and acquiring massive amounts of data can lead to inefficiency in processing time. The lack of interpretability of ML models is also a critical issue, both for end-users and creators. Understanding why the models produce the given output would lead to better applications by identifying biases in the models.

5 Conclusion

In summary, FS-MLU has the potential to revolutionize human-computer communication by enabling conversational agents to interpret and respond to multimodal input. While the research is still in its early stages, the field has seen significant progress with different approaches being proposed. With more standardization of data sets and metrics, we can expect more practical implementations of the models in conversational systems, paving the way to more intuitive human-computer interactions.

References

D. Chen, Z. Huang, X. Wu, S. Ge, and Y. Zou. 2022a. Towards joint intent detection and slot filling via higher-order attention. In *International Joint Conference on Artificial Intelligence*.

D. Chen, Z. Huang, and Y. Zou. 2022b. Leveraging bilinear attention to improve spoken language understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Z. Huang, F. Liu, P. Zhou, and Y. Zou. 2021. Sentiment injected iteratively co-interactive network for spoken language understanding. In *ICASSP*.

Z. Huang, F. Liu, and Y. Zou. 2020. Federated learning for spoken language understanding. In *COLING*.

Z. Huang, M. Rao, A. Raju, Z. Zhang, B. Bui, and C. Lee. 2022. MTL-SLT: multi-task learning for spoken language tasks. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL*.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2017. Stochastic answer networks for machine reading comprehension. *CoRR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

M. Sundararaman, A. Kumar, and J. Vepa. 2021. Phonemebert: Joint language modelling of phoneme sequence and ASR transcript. In *Annual Conference of the International Speech Communication Association*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

P. Zhou, Z. Huang, F. Liu, and Y. Zou. 2020. PIN: A novel parallel interactive network for spoken language understanding. In *International Conference on Pattern Recognition*.