Multimodal language understanding for low-resource languages

Yurii Laba

Ukrainian Catholic University

Abstract. Within this Ph.D. proposal, I've introduced the strategy to enhance numerical data representation for low-resource languages. This is achieved by developing a fine-tuning method that employs pretrained Large Language Model (LLM) for high-resource languages and multimodal data of the specific low-resource language. The initial step involves tailoring this approach for the Ukrainian language. However, the ultimate objective of this research is to establish a comprehensive framework that includes data preprocessing, instantiation, and parameterization of fine-tuning processes. This framework aimed to be flexible across a range of low-resource languages.

Keywords: Multimodal LLM · Natural Language Understanding · Low-Resource Languages.

1 Introduction

Language understanding is a complex and highly relevant problem in the field of natural language processing. Despite significant progress in research and technology development, it has not been adequately solved yet [1]. There are many reasons contributing to this situation, particularly the dynamic nature of languages characterized by constant enrichment with new words and the evolving emotional connotations of existing words. Additionally, some languages remain less studied, leading to limited resources available for their processing.

One of the most promising strategies for addressing the language understanding task is the utilization of LLMs. LLMs have proven to be highly effective methods for solving numerous downstream tasks within the natural language processing domain [2].

Multilingual LLMs present a potent approach to handle languages that are less widely used compared to English. However, for certain specialized tasks, such as Word Sense Disambiguation (WSD), these models may not exhibit optimal performance [3]. While domain adaptation holds promise in enhancing results, a key concern is the adequacy of relevant data for fine-tuning such models.

Despite the cases of successful applications, there are still challenges in directly employing LLMs for multimodal data, including images, audio and text. Through the integration of multimodal input, we can not only broaden the range of tasks that can be effectively addressed but also enhance language comprehension and understanding.

Multimodal models play a vital role in grounding language within real-world contexts, as they combine linguistic information with visual and auditory signals. This integration enables a more profound grasp of language in real-world situations, leading to improved language understanding and performance across various applications, like multimodal question answering, image and video generation, multimodal sentiment analysis, etc [4]. Constructing a multimodal LLM holds the potential to yield more robust numerical representations for the input data. Consequently, this enhanced representation is expected to improve performance across a diverse set of downstream tasks.

The prospect of constructing such a model for a low-resource language like Ukrainian is intriguing. Undoubtedly, accomplishing competitive performance in this endeavor will necessitate the implementation of innovative and sophisticated techniques. Moreover, I perceive advantages not solely in the development of an isolated, finely-tuned model for a particular language, but rather in the entire framework that accepts language-specific data as input and instigates the process of fine-tuning.

I am convinced that the development of such a technique holds promising potential for yielding significant business outcomes. Even a model that enhances WSD performance [3] has captured the curiosity of some business leaders. As a result, some of them have reached out to me regarding the possibility of utilizing the model for practical applications.

The remainder of the proposal is composed of four sections. In the following section, titled Related Works, I outline existing approaches to working with multimodal data and training models based on such data. Moving forward, in the section titled Motivation, I describe my experience in the context of working with low-resource languages. Additionally, I elaborate on what aspects intrigue me in terms of contributing to the development of an approach for building a model using multimodal data for low-resource language. In the subsequent section, titled Approach, I described my own perspective on the development of such an approach. In the final section, Conclusions, I evaluate the potential impact of my contributions in the event of the successful realization of the envisioned concepts.

2 Related Works

As previously explained, improving the accuracy of information representations, also known as embeddings, can be accomplished by creating a common embedding space that can handle visual images and, spoken audio and captions.

One method to achieve this involves utilizing a pair of Convolutional Neural Networks (CNNs). Instead of assigning specific points in the embedding space to entire images and spoken utterances, this technique enables the acquisition of distributed representations that cover both spatial and temporal dimensions. As a result, the models are enabled to directly associate corresponding elements in both visual and auditory modalities, resulting in more efficient and insightful embeddings [5].

However, it is essential to acknowledge that this approach requires an annotated dataset, which can be particularly challenging to obtain for low-resourced languages. Additionally, the reliance on CNNs for processing image and audio data might pose difficulties when attempting transfer learning for application to other languages.

Another intriguing strategy involves utilizing raw signal inputs directly. In this approach, the visual input for the vision modality comprises of 3-channel RGB pixels that depict video frames. The audio input is represented by the amplitude of air density in waveform format, and the textual input is presented as a sequence of words [6] [7]. Moreover, a captivating element of their study involves implementing a semi-supervised methodology to fine tune the LLM. They create sets of video-audio-text triplets and then utilize contrastive learning methods to enhance the model's representations. This strategy shows great potential for transfer learning and future semi-supervised fine tuning, especially when dealing with languages that have limited resources.

An alternative approach to harnessing multiple modalities involves employing diverse data types or formats for model training. This encompasses monomodal data, such as a text corpus; cross-modal paired data, like image-caption pairs; and interleaved multimodal data, where documents incorporate both images and text in arbitrary sequences [8].

Initiating the training procedure, the model first utilizes monomodal data to facilitate representation learning. This entails language modeling with textual data, whereby the model undergoes pretraining on tasks involving tasks such as instruction comprehension, contextual learning, and other language-centric challenges. This initial phase equips the model with a robust grasp of textual information and linguistic structures. During this process, the models are primarily engaged in the task of predicting the subsequent token. They learn to generate the next token in a sequence based on contextual cues provided by preceding tokens. The model trained in this manner presents an intriguing prospect for subsequent fine-tuning in the context of low-resource languages.

3 Motivation

Over the previous year, my research efforts have concentrated on the language understanding field, with a particular emphasis on the low-resource language. Specifically, my work have been centered around addressing the quite challenging task of Word Sense Disambiguation (WSD) for the Ukrainian language, improving language understanding [3].

In the course of this study, as a foundational step, we developed an approach for generating a dataset for the assessment of WSD in the Ukrainian language. This dataset served as a valuable means for evaluating the efficacy of our model during the experimentation phase. Additionally, we formulated a technique to construct a dataset for semi-supervised learning based on UberText 2.0 [15]. By utilizing such dataset we were able to engineer and fine-tune the LLM using contrastive learning approach.

By employing our proposed methodology, we achieved notable enhancements in the quality of contextual embeddings for words with multiple meanings. This, in turn, led to a substantial enhancement in the performance of Word Sense Disambiguation within the Ukrainian language context.

Furthermore, we have constructed an all-encompassing framework that facilitates the utilization of diverse prediction and pooling methodologies, LLMs, and the generation of numerous performance assessments, as well as generation of evaluation and training datasets. This framework holds the potential to be applied in addressing the WSD challenge across different languages.

Considering the subject of my previous research, I have chosen to expand the scope slightly and embark on an exploration of language understanding, employing multimodal data as an integral component of my experimentation. The objective is to investigate the potential of leveraging pre-trained multimodal LLMs from other languages, especially high resource languages, to facilitate the training of the multimodal LLM for the low resource language.

The primary stage in the adaptation of a LLM often encompasses the fine-tuning of hyperparameters. Under specific circumstances, this strategy can indeed yield positive outcomes [9] [10]. Nonetheless, it becomes apparent that the effectiveness of altering hyperparameters could be limited if the selected model architecture is not well-suited to the unique linguistic characteristics of the low-resource language.

Another instinctive strategy involves utilizing an unlabeled corpus of low-resource language or corpus from some target domain for fine tuning LLM [11]. Nonetheless, the efficacy of these techniques cannot be assured in scenarios where the size of the unlabeled corpora within the low-resource language is limited. In this scenario, the utilization of multimodal data could potentially offer valuable insights, as it introduces multiple data sources that have the capacity to acquire a broader spectrum of information.

Fine-tuning of the LLM frequently draws upon Masked Language Modeling (MLM) [12]. However, contrastive loss has exhibited heightened efficacy compared to MLM. This is attributed to its capability to encapsulate more intricate semantic relationships and contextual nuances present in the data [13]. While both approaches hold prominence within the realm of self-supervised learning, contrastive loss presents notable merits for specific tasks and situations. Notably, contrastive learning yields embeddings of superior quality, proving especially beneficial for downstream applications [14].

In light of the considerations above, I am of the opinion that building an approach to utilize multimodal data alongside pre-trained LLM may constitute a viable approach to enhancing the numerical representation of input data, especially in the context of low-resource languages.

4 Approach

The iterative methodology for approaching my preliminary experiments encompasses the subsequent stages:

- *Developing a pipeline to collect a semi supervised multimodal dataset for low-resource language*
  I will commence my experiments by assembling a training dataset for the Ukrainian language. The underlying concept revolves around curating such a dataset through the utilization of the Ukrainian segment of YouTube. Access to videos can be obtained via API. From these videos, we can extract both audio and visual modalities. It's important to note that not all YouTube channels provide manually generated subtitles. To acquire subtitles in such cases, we can employ automatic speech recognition methods.
- *Preprocessing of the collected semi supervised dataset*
  Upon assembling a multimodal dataset, I will proceed to assess its quality and explore potential enhancements. One of the primary challenges that may arise involves a discrepancy between audio and subtitles. To investigate this, I will cut the audio-containing speech into fragments of specific durations, aligning them with corresponding subtitles by timestamps. Subsequently, a manual analysis of these aligned data points will be conducted. Additionally, I will employ Automatic Speech Recognition (ASR) to transcribe the audio content, enabling the calculation of the Word Error Rate (WER) for each audio fragment. This analytical approach will facilitate the identification of areas where the model exhibits substantial errors, allowing for a comprehensive investigation into the underlying causes. An avenue I perceive for enhancing the alignment of audio and text involves the utilization of a forced alignment approach [16].
- *Developing a framework for multimodal domain adaptation*
  As a preliminary step, my intent is to employ the VATT model [6]. This model operates under the Apache license, which expressly permits its utilization for research purposes. The model checkpoints are available on GitHub. Subsequently, I plan to execute the baseline model based on the subset of, for example, Spoken Language Understanding Evaluation (SLUE) benchmark [17] during fine tuning. I also would like to utilize some custom evaluation tasks like WSD [3]. This will pave the way for the subsequent phase of refining the model through fine-tuning. In this context, my focus shifts to the intricate process of selecting and harmonizing the loss function. I anticipate delving into various advanced techniques, including but not limited to Layer-wise Learning Rate Decay (LLRD), Stochastic Weight Averaging (SWA), among others. Furthermore, during this stage, I envisage engaging in the exploration of potential enhancements for approaches related to the alignment of embeddings across video, audio, and textual modalities.
- *Establishment a pipeline for parameter logging and tracking of training metrics*
  At this juncture, my aim is to configure all the pertinent metrics that I intend to monitor throughout the model's fine-tuning process. During this iterative endeavor, a scenario might arise wherein the loss function demonstrates a downward trend, yet the evaluation metrics could exhibit a degradation in performance.

- *Developing an computational effective training pipeline*
  In this phase, my objective is to establish an efficient model fine-tuning process that capitalizes on the utilization of multiple GPUs for enhanced computational capabilities.
- *Evaluation and analysis of the outcomes achieved*
  At this juncture, I intend to conduct thorough model testing, encompassing not only subsets but also full-scale versions of the datasets. This comprehensive evaluation aims to elucidate the effectiveness of the approach/experimentation under consideration.

In the subsequent phase, my vision involves delving into an exploration aimed at integrating the proposed approach into a comprehensive framework. The objective is to facilitate its broader application to additional low-resource languages.

5 Conclusions

In the present Ph.D. proposal, I have outlined an approach to enhance input data's numerical representation by developing a fine-tuning approach utilizing LLM and multimodal data. The preliminary phase of this study involves the training of the model specifically for the Ukrainian language. However, the ultimate outcome of this research endeavor is the development of a comprehensive framework that facilitates the data preprocessing, instantiation, and parameterization of fine-tuning procedures for diverse low-resource languages.

I perceive substantial value in this research, both within the business sector and the academic community. Primarily, suppose the envisioned goals are successfully realized. In that case, this approach holds the potential to emerge as a state-of-the-art methodology for various downstream tasks, particularly among low-resource languages. Nonetheless, the significance of my work extends to the business realm as well, as enterprises actively leverage language models to engage with their clients.

References

1. Lenci, Alessandro. "Understanding Natural Language Understanding Systems. A Critical Analysis." arXiv preprint arXiv:2303.04229 (2023).
2. Yang, Jingfeng, et al. "Harnessing the power of llms in practice: A survey on chatgpt and beyond." arXiv preprint arXiv:2304.13712 (2023).
3. Laba, Yurii, et al. "Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation." Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). 2023.
4. Yin, Shukang, et al. "A Survey on Multimodal Large Language Models." arXiv preprint arXiv:2306.13549 (2023).
5. Harwath, David, et al. "Jointly discovering visual objects and spoken words from raw sensory input." Proceedings of the European conference on computer vision (ECCV). 2018.

6. Akbari, Hassan, et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text." Advances in Neural Information Processing Systems 34 (2021): 24206-24221.
7. Alayrac, Jean-Baptiste, et al. "Self-supervised multimodal versatile networks." Advances in Neural Information Processing Systems 33 (2020): 25-37.
8. Huang, Shaohan, et al. "Language is not all you need: Aligning perception with language models." arXiv preprint arXiv:2302.14045 (2023).
9. Araabi, Ali, and Christof Monz. "Optimizing transformer for low-resource neural machine translation." arXiv preprint arXiv:2011.02266 (2020).
10. Lankford, Séamus, Haithem Afli, and Andy Way. "Transformers for low-resource languages: is féidir linn!." (2021).
11. Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.
12. Chalkidis, Ilias, et al. "LEGAL-BERT: The muppets straight out of law school." arXiv preprint arXiv:2010.02559 (2020).
13. Fu, Zhiyi, et al. "Contextual representation learning beyond masked language modeling." arXiv preprint arXiv:2204.04163 (2022).
14. Hu, Xiyang, et al. "Language Agnostic Multilingual Information Retrieval with Contrastive Learning." arXiv preprint arXiv:2210.06633 (2022).
15. Chaplynskyi, Dmytro. "Introducing UberText 2.0: a corpus of modern Ukrainian at scale." Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). 2023.
16. Montreal Forced Aligner, https://montreal-forced-aligner.readthedocs.io/en/latest/index.html.
17. Shon, Suwon, et al. "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.