MML 2022

The 1st Workshop on Multilingual Multimodal Learning

Proceedings of the Workshop

May 27, 2022

The MML organizers gratefully acknowledge the support from the following sponsors.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-44-5

Preface

Welcome to the Workshop on Multilingual Multimodal Learning (MML)! Multilingual multimodal NLP, which presents new and unique challenges. Multilingual multimodal NLP is one of the areas that suffer the most from language imbalance issues. Texts in most multimodal datasets are usually only available in high-resource languages. Further, multilingual multimodal research provides opportunities to investigate culture-related phenomena. On top of the language imbalance issue in text-based corpora and models, the data of additional modalities (e.g. images or videos) are mostly collected from North American and Western European sources (and their worldviews). As a result, multimodal models do not capture our world's multicultural diversity and do not generalise to out-of-distribution data from minority cultures. The interplay of the two issues leads to extremely poor performance of multilingual multimodal systems in real-life scenarios. This workshop offers a forum for sharing research efforts towards more inclusive multimodal technologies and tools to assess them.

This volume includes the 3 papers presented at the workshop. We received a batch of high-quality research papers, and decided to finally accept 3 out of 6 fully reviewed submissions. MML 2022 was co-located with the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) and was held on May 27, 2022 as a hybrid workshop.

It is great to see the accepted papers discussing some of the most important topics for MML: evaluating and developing multimodal models for different languages and in low-resource settings.

The first edition of the MML workshop also hosted a shared task on multilingual visually grounded reasoning. The task was centered around the MaRVL dataset. This dataset extends the NLVR2 task to multicultural and multilingual (Indonesian, Mandarin, Swahili, Tamil, Turkish) inputs: Given two images and a textual description, a system needs to predict whether the description applies to both images (True/False). The shared task is developed to encourage new methods for multilingual multimodal models, with the restriction that models should be publicly available or trained on publicly available data to encourage open source.

We take this opportunity to thank the MML program committee for their help and thorough reviews. We also thank the authors who presented their work at MML, and the workshop participants for the valuable feedback and discussions. Finally, we are deeply honored to have four excellent talks from our invited speakers: David Ifeoluwa Adelani, Lisa-Anne Hendricks, Lei Ji, and Preethi Jyothi.

The MML 2022 workshop organizers,

Emanuele Bugliarello, Kai-Wei Chang, Desmond Elliott, Spandana Gella, Aishwarya Kamath, Liunian Harold Li, Fangyu Liu, Jonas Pfeiffer, Edoardo M. Ponti, Krishna Srinivasan, Ivan Vulić, Yinfei Yang, Da Yin

Organizing Committee

Workshop Chairs

Emanuele Bugliarello, University of Copenhagen
Kai-Wei Chang, UCLA
Desmond Elliott, University of Copenhagen
Spandana Gella, Amazon Alexa AI
Aishwarya Kamath, NYU
Liunian Harold Li, UCLA
Fangyu Liu, University of Cambridge
Jonas Pfeiffer, TU Darmstadt
Edoardo Maria Ponti, MILA Montreal
Krishna Srinivasan, Google Research
Ivan Vulić, University of Cambridge & PolyAI
Yinfei Yang, Google Research
Da Yin, UCLA

Invited Speakers

David Ifeoluwa Adelani, Saarland University
Lisa Anne Hendricks, DeepMind
Lei Ji, Microsoft Research Asia
Preethi Jyothi, IIT Bombay

Program Committee

Program Committee

Arjun Reddy Akula, UCLA
Benno Krojer, McGill University
Chen Cecilia Liu, TU Darmstadt
Duygu Ataman, New York University
Constanza Fierro, University of Copenhagen
Gregor Geigle, TU Darmstadt
Hao Tan, Adobe Systems
Kuan-Hao Huang, UCLA
Laura Cabello Piqueras, University of Copenhagen
Rongtian Ye, Aalto University
Rémi Lebret, EPFL
Sebastian Schuster, New York University
Mandy Guo, Cornell University
Zarana Parekh, Carnegie Mellon University

Invited Talk: Multilingual Language Model Adaptive Fine-Tuning: A Study on African Languages

David Ifeoluwa Adelaini
Saarland University

Abstract: Multilingual pre-trained language models (PLMs) have demonstrated impressive performance on several downstream tasks on both high resourced and low-resourced languages. However, there is still a large performance drop for languages unseen during pre-training, especially African languages. One of the most effective approaches to adapt to a new language is language adaptive fine-tuning (LAFT) — fine-tuning a multilingual PLM on monolingual texts of a language using the pre-training objective. However, African languages with large monolingual texts are few, and adapting to each of them individually takes large disk space and limits the cross-lingual transfer abilities of the resulting models because they have been specialized for a single language. As an alternative, we adapt PLM on several languages by performing multilingual adaptive fine-tuning (MAFT) on 17 most-resourced African languages and three other high-resource languages widely spoken on the continent – English, French, and Arabic to encourage cross-lingual transfer learning. Additionally, to further specialize the multilingual PLM, we removed vocabulary tokens from the embedding layer that corresponds to non-African writing scripts before MAFT, thus reducing the model size by 50%. Our evaluation on two multilingual PLMs (AfriBERTa and XLM-R) and three NLP tasks (NER, news topic classification, and sentiment classification) shows that our approach is competitive to applying LAFT on individual languages while requiring significantly less disk space.

Bio: David Ifeoluwa Adelani is a doctoral student in computer science at Saarland University, Saarbrücken, Germany, and an active member of Masakhane NLP - a grassroots organization whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans. His current research focuses on NLP for African languages, multilingual representation learning, and privacy in NLP.

Invited Talk: Multimodal Video Understanding with Language Guidance

Lei Ji
Microsoft Research Asia

Abstract: Video naturally comprises of multiple modalities including visual content, language (speech text or meta data), as well as audio. Language inside video provides important semantic guidance for multimodal video understanding. On the one hand, language can be taken as side information to enhance video information with fusion mechanism. On the other hand, language can be used as semantic supervision for video representation learning with self-supervised techniques. Multimodal video-language pretraining models trained on a large-scale dataset are effective for multimodal understanding tasks including vision language matching, captioning, sentiment analysis as well multilingual multimodal tasks as a step forward.

Bio: Lei Ji is a senior researcher at the Natural Language Computing group of Microsoft Research Asia. Her research focuses are vision and language multimodal learning, pretraining, and knowledge mining and reasoning. She has published papers at the top-tier conferences including ACL, AAAI, IJCAI, ACMMM, KDD, CIKM, patents, and transferred these innovative techniques to Microsoft products as well as other external partners.

Invited Talk: Digging Deeper into Multimodal Transformers

Lisa Anne Hendricks
DeepMind

Abstract: Multimodal transformers have had great success on a wide variety of multimodal tasks. This talk will consider what factors contribute to their success as well as what still proves challenging for these models. I will first consider how the choice of training dataset, architecture, and loss function contribute to multimodal transformer performance on a zero-shot image retrieval task. Next, using the newly collected SVO-Probes dataset, I will demonstrate that fine-grained verb understanding is challenging for multimodal transformers and offers an interesting testbed to study multimodal understanding.

Bio: Lisa Anne Hendricks is a research scientist on the Language Team at DeepMind. She received her PhD from Berkeley in May 2019, and a BSEE (Bachelor or Science in Electrical Engineering) from Rice University in 2013. Her research focuses on the intersection of language and vision. She is particularly interested in analyzing why models work, explainability, and mitigating/measuring bias in AI models.

Invited Talk: New Challenges in Learning with Multilingual and Multimodal Data

Preethi Jyothi
IIT Bombay

Abstract: During communication, humans can naturally combine their knowledge about different languages and process simultaneous cues from different modalities. Even as machine learning has made great strides in natural language processing, problems related to multilinguality and multimodality remain largely unsolved. In recent years, each of these issues has received significant attention from the research community. In this talk, we will discuss some of our work on both multilinguality and multimodality, specifically code-switching and audio-visual learning, respectively. Further, towards understanding the additional difficulties that arise when multilinguality and multimodality are present together, we will also describe our recent work on a new multimodal multilingual dataset in Indian languages.

Bio: Preethi Jyothi is an Assistant Professor in the Department of Computer Science and Engineering at IIT Bombay. Her research interests are broadly in machine learning applied to speech and language, specifically focusing on Indian languages and low-resource settings. She was a Beckman Postdoctoral Fellow at the University of Illinois at Urbana-Champaign from 2013 to 2016. She received her Ph.D. from The Ohio State University in 2013. Her doctoral thesis dealt with statistical models of pronunciation in conversational speech and her work on this topic received a Best Student Paper award at Interspeech 2012. She co-organised a research project on probabilistic transcriptions at the 2015 Jelinek Summer Workshop on Speech and Language Technology, for which her team received a Speech and Language Processing Student Paper Award at ICASSP 2016. She was awarded a Google Faculty Research Award in 2017 for research on accented speech recognition. She currently serves on the ISCA SIGML board and is a member of the Editorial Board of Computer Speech and Language.

Table of Contents

Program

Friday, May 27, 2022

09:20 - 09:30 *Opening Remarks*

09:30 - 10:30 *Invited Talk 1: David Ifeoluwa Adelaini*

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Invited Talk 2: Lei Ji*

12:00 - 12:30 *Findings from the MaRVL Shared Task*

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Invited Talk 3: Lisa Anne Hendricks*

15:00 - 15:45 *Workshop Papers: Archival and Non-Archival*

15:45 - 16:00 *Short Break*

16:00 - 17:00 *Invited Talk 4: Preethi Jyothi*

17:00 - 17:10 *Concluding Remarks*