Tutorial on Multimodal Machine Learning

Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh
Carnegie Mellon University
{morency,pliang,abagherz}@cs.cmu.edu
https://cmu-multicomp-lab.github.io/mmml-tutorial/naacl2022/

Abstract

Multimodal machine learning involves integrating and modeling information from multiple heterogeneous and interconnected sources of data. It is a challenging yet crucial area with numerous real-world applications in multimedia, affective computing, robotics, finance, HCI, and healthcare. This tutorial, building upon a new edition of a survey paper on multimodal ML as well as previously-given tutorials and academic courses, will describe an updated taxonomy on multimodal machine learning synthesizing its core technical challenges and major directions for future research.

1 Introduction

Multimodal machine learning is a vibrant multi-disciplinary research field that addresses some original goals of AI by integrating and modeling multiple communicative modalities, including linguistic, acoustic, and visual messages. With the initial research on audio-visual speech recognition and more recently with language & vision projects such as image and video captioning, visual question answering, and language-guided reinforcement learning, this research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities.

This tutorial builds upon the annual course on Multimodal Machine Learning taught at Carnegie Mellon University and is a revised version of the previous tutorials on multimodal learning at CVPR 2021, ACL 2017, CVPR 2016, and ICMI 2016. These previous tutorials were based on our earlier survey on multimodal machine learning, which introduced an initial taxonomy for core multimodal challenges (Baltrusaitis et al., 2019). The present tutorial is based on a revamped taxonomy of the core technical challenges and updated concepts about recent work in multimodal machine learning (Liang et al., 2022). The tutorial will be centered around six core challenges in multimodal machine learning:

1. Representation: A first fundamental challenge is to learn representations that exploit cross-modal interactions between individual elements of different modalities. The heterogeneity of multimodal data makes it particularly challenging to learn multimodal representations. We will cover fundamental approaches for (1) *representation fusion* (integrating information from 2 or more modalities, effectively reducing the number of separate representations), (2) *representation coordination* (interchanging cross-modal information with the goal of keeping the same number of representations but improving multimodal contextualization), and (3) *representation fission* (creating a new disjoint set of representations, usually larger number than the input set, that reflects knowledge about internal structure such as data clustering or factorization).

2. Alignment: A second challenge is to identify the connections between all elements of different modalities using their structure and cross-modal interactions. For example, when analyzing the speech and gestures of a human subject, how can we align specific gestures with spoken words or utterances? Alignment between modalities is challenging since it may exist at different (1) *granularities* (words, utterances, frames, videos), involve varying (2) *correspondences* (one-to-one, many-to-many, or not exist at all), and depend on long-range (3) *dependencies*.

3. Reasoning is defined as composing knowledge from multimodal evidences, usually through multiple inferential steps, to exploit multimodal alignment and problem structure for a specific task. This relationship often follows some hierarchical structure, where more abstract concepts are defined higher in the hierarchy as a function of less abstract concepts. Multimodal reasoning involves the subchallenges of capturing this (1) *structure* (through domain knowledge or discovered from

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics
Human Language Technologies: Tutorial Abstracts, pages 33 - 38
July 10-15, 2022 ©2022 Association for Computational Linguistics

data), the parameterization of (2) *concepts* (dense vs interpretable and symbolic), and the type of (3) *composition* (simple vs complex relationships).

4. Generation: The fourth challenge involves learning a generative process to produce raw modalities that reflect cross-modal interactions, structure and coherence. We categorize its subchallenges into (1) *summarization* (summarizing multimodal data to reduce information content while highlighting the most salient parts of the input), (2) *translation* (translating from one modality to another and keeping information content while being consistent with cross-modal interactions), and (3) *creation* (simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities). We will also cover advances in evaluation and ethical concerns around generated content.

5. Transference: A fifth challenge is to transfer knowledge between modalities and their representations, usually to help the target modality, which may be noisy or with limited resources. Exemplified by algorithms of (1) *transfer* (fine-tuning pre-trained models for a downstream task involving the target modality), (2) *representation enrichment* (transfer through a joint model sharing representation spaces between both modalities), and (3) *model induction* (keeping individual unimodal models separate but transferring information across these models), how can knowledge learned from one modality (e.g., predicted labels or representation) help a computational model trained on a different modality?

6. Quantification involves a deeper measurement and theoretical study of multimodal models to better understand their (1) *output qualities* (the extent to which models are predictive, efficient, and robust under natural and targeted modality imperfections), (2) *internal mechanics* (understanding the internal modeling of multimodal information and cross-modal interactions), and (3) *modality tradeoffs* (quantifying the utility and risks of each input modality, while balancing these tradeoffs for reliable real-world usage). It is important to obtain a deeper understanding of the data, modeling, and optimization challenges involved when learning from heterogeneous data in order to improve their robustness, interpretability, and reliability in real-world multimodal applications.

Type of tutorial: This tutorial will begin with basic concepts related to multimodal research before describing cutting-edge research in the context of the six core challenges.

Target audience and expected background: We expect the audience to have an introductory background in machine learning and deep learning, including a basic familiarity of commonly-used unimodal building blocks such as convolutional, recurrent, and self-attention models.

2 Tutorial outline

This tutorial will be a revised edition of our previously-organized tutorials at CVPR 2022, CVPR 2021, ACL 2017, CVPR 2016, and ICMI 2016 which were roughly 3-4 hours long. This revision defines a new iteration of the taxonomy that has been updated to help researchers tackle modern multimodal challenges. The tutorial outline is shown below:

Introduction (30 mins)

- What is Multimodal? Definitions, dimensions of heterogeneity and cross-modal interactions.
- Historical view and multimodal research tasks.
- Core technical challenges: representation, alignment, transference, reasoning, generation, and quantification.
- Unimodal language, visual, and acoustic representations.

Representation (30 mins)

- Representation fusion: fusion strategies, multimodal auto-encoder.
- Representation coordination: contrastive learning, vector-space models, canonical correlation analysis.
- Representation fission: factorization, component analysis, disentanglement.

===== BREAK =====
Alignment (25 mins)

- Granularity: segmentation, clustering, unit definition.

- Correspondences: latent alignment approaches, attention models, multimodal transformers, multi-instance learning.
- Dependency types: Attention models, graph neural networks, multimodal transformers, multi-instance learning.

Transference (25 mins)

- Modality transfer: losses, hallucination, cross-modal transfer.
- Foundation models: pre-trained models and adaptation.
- Model induction: co-training, cross-modal learning.

===== BREAK =====

Reasoning (20 mins)

- Structure: hierarchical, graphical, temporal, and interactive structure, structure discovery.
- Concepts: dense and neuro-symbolic.
- Composition: causal and logical relationships.
- Knowledge: external knowledge bases, commonsense reasoning.

Generation (15 mins)

- Summarization, translation, and creation.
- Model evaluation and ethical concerns.

Quantification (25 mins)

- Output qualities: generalization, robustness, complexity.
- Internal mechanics: interpretability, understanding cross-model interactions.
- Modality tradeoffs: dataset biases, social biases, theoretical benefits, optimization challenges.

Future directions and conclusion (10 mins)

3 Tutorial details

Included work: The tutorial is based on an updated version (Liang et al., 2022) of the broadly cited survey on multimodal ML (Baltrusaitis et al., 2019) which covers fundamental work in multimodal, including affective computing (Poria et al., 2017), audio-visual learning, image and video-based question answering (Agrawal et al., 2017), media description (Vinyals et al., 2016), multimodal machine translation (Yao and Wan, 2020), multimodal reinforcement learning (Luketina et al., 2019), and social impacts of real-world multimodal learning (Liang et al., 2021). The updated survey will be released with this tutorial, following the six core challenges mentioned earlier. While the taxonomy is developed by the organizers, most of the presented work comes from the broader research community.

Diversity: This tutorial will cover multilingual tasks (e.g. multimodal machine translation) and multiple research domains (image, text, audio). This tutorial brings together faculty, graduate students, and postdoctoral researchers. Slides will also be dedicated to low-data language and modality scenarios.

Reading list: We suggest the following reading list. These papers can be skimmed through before the tutorial, and are also well-served as reading material for after the tutorial. A more comprehensive reading list can be found in the multimodal ML courses at CMU, see `https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/` and `https://cmu-multicomp-lab.github.io/mmml-course/fall2020/` for more details.

1. General: Fundamentals of Multimodal Machine Learning: A Taxonomy and Open Challenges (Liang et al., 2022)
2. General: Multimodal Machine Learning: A Survey and Taxonomy (Baltrusaitis et al., 2019)
3. General: Representation learning: A review and new perspectives (Bengio et al., 2013)
4. Representation: Multiplicative Interactions and Where to Find Them (Jayakumar et al., 2020)

5. Representation: Multimodal Learning with Deep Boltzmann Machines (Srivastava and Salakhutdinov, 2014)

6. Representation: Learning Factorized Multimodal Representations (Tsai et al., 2019)

7. Alignment: Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers (Hendricks et al., 2021)

8. Alignment: Deep canonical correlation analysis (Andrew et al., 2013)

9. Transference: Vokenization: Improving Language Understanding via Contextualized, Visually-Grounded Supervision (Tan and Bansal, 2020)

10. Transference: Foundations of Multimodal Co-learning (Zadeh et al., 2020)

11. Reasoning: The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision (Mao et al., 2018)

12. Reasoning: A Survey of Reinforcement Learning Informed by Natural Language (Luketina et al., 2019)

13. Reasoning: VQA-LOL: Visual Question Answering Under the Lens of Logic (Gokhale et al., 2020)

14. Generation: Cross-modal Coherence Modeling for Caption Generation (Alikhani et al., 2020)

15. Generation: Zero-shot Text-to-Image Generation (Ramesh et al., 2021)

16. Quantification: MultiBench: Multiscale Benchmarks for Multimodal Representation Learning (Liang et al., 2021)

17. Quantification: M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis (Wang et al., 2021)

18. Quantification: Women also Snowboard: Overcoming Bias in Captioning Models (Hendricks et al., 2018)

4 Organizers

Louis-Philippe Morency (LTI, CMU) is an Associate Professor at CMU Language Technology Institute where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He received his Ph.D. and Master's degrees from MIT Computer Science and Artificial Intelligence Laboratory. In 2008, Dr. Morency was selected as one of "AI's 10 to Watch" by IEEE Intelligent Systems. He has received 7 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion, and computational models of human communication dynamics. He has taught 10 editions of the multimodal machine learning course at CMU and before that at the University of Southern California. He has given multiple tutorials on this topic, including at ACL 2017, CVPR 2016, and ICMI 2016.

Paul Pu Liang (MLD, CMU) is a Ph.D. student in Machine Learning at Carnegie Mellon University, advised by Louis-Philippe Morency and Ruslan Salakhutdinov. His research is centered around building socially intelligent embodied agents with the ability to perceive and engage in multimodal human communication. He was a recipient of the distinguished student paper award at the NeurIPS 2019 workshop on federated learning and the best paper honorable mention award at ICMI 2017. He organized the workshop on human multimodal language at ACL 2020 and ACL 2018, the workshop on tensor networks at NeurIPS 2020, and was a workflow chair for ICML 2019.

Amir Zadeh (LTI, CMU) is a Postdoctoral Associate at Carnegie Mellon University. Prior to that, he received his Ph.D. from Language Technologies Institute, School of Computer Science, Carnegie Mellon University. His work is focused on multimodal learning, especially modeling multimodal language. He is the creator of several resources in this area including CMU-MOSEAS, CMU-MOSEI, and CMU-MOSI datasets. He organized the 1st and 2nd Workshop and Grand-Challenge on Multimodal Language in ACL 2018 and ACL 2020 respectively. His work has been published in ACL, EMNLP, NAACL, CVPR, and ICLR conferences.

5 Logistics

Audience size and previous editions: Our tutorial build upon 5 previous tutorials:

- CVPR 2022: 100-150 attendees. 6-hour tutorial https://cmu-multicomp-lab.github.io/mmml-tutorial/cvpr2022/
- CVPR 2021: 100-150 attendees. 6-hour tutorial https://audio-visual-scene-understanding.github.io/
- ACL 2017: 100-150 attendees. 4-hour tutorial (Morency and Baltrušaitis, 2017)
- CVPR 2016: 150-200 attendees. 4-hour tutorial: https://sites.google.com/site/multiml2016cvpr/
- ICMI 2016: 50-60 attendees, 3-hour tutorial: https://icmi.acm.org/2016/index.php?id=tutorial

This tutorial builds upon the annual Multimodal Machine Learning course taught at CMU (course 11-877 and 11-777). For recent iterations of the course, the materials are publicly available at:

- https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/
- https://cmu-multicomp-lab.github.io/mmml-course/fall2020/
- https://piazza.com/cmu/fall2019/11777/resources

In Fall 2020, the course was virtual and all lecture videos were recorded and publicly available on YouTube. These videos have become hugely popular, amassing close to $50,000$ views.

Ethics statement: Multimodal models used in real-world applications can pose several considerations such as having higher time and space complexity as compared to unimodal tasks, privacy and security resulting from human-centric multimodal data, and capturing social biases through human language, human faces, human audio, and other multimodal data sources. Our tutorial will cover these risks of multimodal learning and describe recent work towards addressing these critical issues.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual question answering. *International Journal of Computer Vision*.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*.

Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. 2020. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. In *NeurIPS Datasets and Benchmarks Track*.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations of multimodal machine learning: A taxonomy and open challenges. *arXiv preprint*.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *IJCAI*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2018. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*.

Louis-Philippe Morency and Tadas Baltrušaitis. 2017. Multimodal machine learning: Integrating language, vision and speech. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 3–5, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal learning with deep boltzmann machines. *J. Mach. Learn. Res.*, 15(1):2949–2980.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. *ICLR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. 2021. M2lens: visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. 2020. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193.