MULTIZOO & MULTIBENCH: A Standardized Toolkit for Multimodal Deep Learning

Paul Pu Liang, Yiwei Lyu, Xiang Fan, {PLIANG,YLYU1,XIANGFAN}@CS.CMU.EDU
Arav Agarwal, Yun Cheng, {ARAVA,YUNCHENG}@CS.CMU.EDU
Louis-Philippe Morency, Ruslan Salakhutdinov {MORENCY,RSALAKHU}@CS.CMU.EDU
Machine Learning Department and Language Technologies Institute, Carnegie Mellon University

Editor: Antti Honkela

Abstract

Learning multimodal representations involves integrating information from multiple heterogeneous sources of data. In order to accelerate progress towards understudied modalities and tasks while ensuring real-world robustness, we release MULTIZOO, a public toolkit consisting of standardized implementations of > 20 core multimodal algorithms and MULTIBENCH, a large-scale benchmark spanning 15 datasets, 10 modalities, 20 prediction tasks, and 6 research areas. Together, these provide an automated end-to-end machine learning pipeline that simplifies and standardizes data loading, experimental setup, and model evaluation. To enable holistic evaluation, we offer a comprehensive methodology to assess (1) generalization, (2) time and space complexity, and (3) modality robustness. MULTIBENCH paves the way towards a better understanding of the capabilities and limitations of multimodal models, while ensuring ease of use, accessibility, and reproducibility. Our toolkits are publicly available, will be regularly updated, and welcome inputs from the community[1].

Code: `https://github.com/pliang279/MultiBench`

Documentation: `https://multibench.readthedocs.io/en/latest/`

Keywords: Multimodal learning, Representation learning, Benchmarks, Open Source Software

1. Introduction

The research field of multimodal machine learning (ML) brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources (Baltrušaitis et al., 2018; Liang et al., 2022). At its core lies the learning of *multimodal representations* that capture correspondences between modalities for prediction, and has emerged as a vibrant interdisciplinary field of immense importance and with extraordinary potential in multimedia (Naphade et al., 2006; Liang et al., 2023), affective computing (Liang et al., 2019; Poria et al., 2017), robotics (Kirchner et al., 2019; Lee et al., 2019), finance (Hollerer et al., 2018), dialogue (Pittermann et al., 2010), human-computer interaction (Dumas et al., 2009; Obrenovic and Starcevic, 2004), and healthcare (Frantzidis et al., 2010; Xu et al., 2019). In order to accelerate research in building general-purpose multimodal models across diverse research areas, modalities, and tasks, we contribute MULTIBENCH (Figure 1), a systematic and unified large-scale benchmark that brings us closer to the requirements of real-world multimodal applications. MULTIBENCH contains a diverse set of 15 datasets spanning 10 modalities and testing for 20 prediction tasks across 6 distinct research areas, and is designed to comprehensively evaluate generalization across domains and modalities, complexity during training and inference, and robustness to noisy and missing modalities. Additionally, we release MULTIZOO, a public toolkit consisting of standardized implementations of > 20 core multimodal algorithms in a modular fashion

[1] MULTIBENCH was previously published at NeurIPS 2021 (Liang et al., 2021), although the datasets and algorithms were the central contributions of that publication, not the software. This paper focuses on the open-source software along with a larger collection of datasets, algorithms, and evaluation metrics.

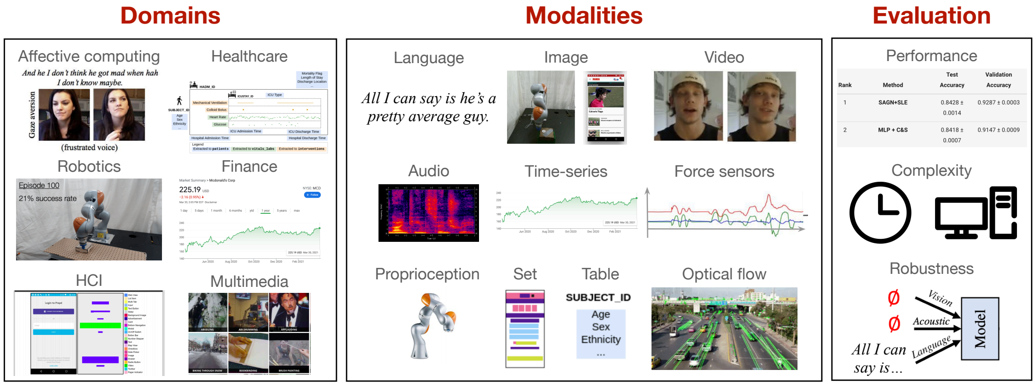

Figure 1: MULTIBENCH contains a diverse set of 15 datasets spanning 10 modalities and testing for more than 20 prediction tasks across 6 distinct research areas, and enables standardized, reliable, and reproducible large-scale benchmarking of multimodal models for performance, complexity, and robustness.

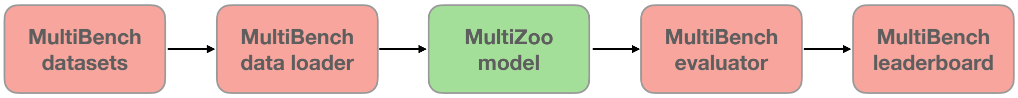

Figure 2: Our MULTIBENCH toolkit provides a machine learning pipeline across data processing, data loading, multimodal models, evaluation metrics, and a public leaderboard to encourage accessible, standardized, and reproducible research in multimodal representation learning.

to enable accessibility for new researchers, compositionality of approaches, and reproducibility of results. Together, these public resources ensure ease of use, accessibility, and reproducibility, and they will be continually expanded in courses, workshops, and competitions around the world.

2. MULTIBENCH and MULTIZOO

MULTIBENCH provides a standardized machine learning pipeline that starts from data loading to running multimodal models, providing evaluation metrics, and a public leaderboard to encourage future research in multimodal representation learning (see Figure 2).

MULTIBENCH datasets: Table 1 shows an overview of the datasets provided in MULTIBENCH, which span research areas in multimedia, affective computing, robotics, finance, human-computer interaction, and healthcare, more than 15 datasets, 10 modalities, and 20 prediction tasks.

MULTIZOO: A zoo of multimodal algorithms: To complement MULTIBENCH, we release a comprehensive toolkit, MULTIZOO, as starter code for multimodal algorithms which implements 20 methods spanning different methodological innovations in (1) data preprocessing, (2) fusion paradigms, (3) optimization objectives, and (4) training procedures (see Figure 3). Each of these algorithms are chosen because they provide unique perspectives to the technical challenges in multimodal learning (Baltrušaitis et al., 2018) (see Table 2 for details).

Evaluation protocol: MULTIBENCH contains evaluation scripts for the following holistic desiderata in multimodal learning: (1) Performance: We standardize evaluation using MSE and MAE for regression, as well as accuracy, micro & macro F1-score, and AUPRC for classification. (2) Complexity: We record the amount of information taken in bits (i.e., data size), the number of model parameters, as well as time and memory resources required during the entire training process. Real-world models may also need to be small and compact to run on mobile devices (Radu et al., 2016) so we also report inference time and memory on CPU and GPU. The datasets and models

Table 1: MULTIBENCH provides a comprehensive suite of 15 datasets covering a diverse range of 6 research areas, dataset sizes, 10 input modalities (in the form of ℓ: language, i: image, v: video, a: audio, t: time-series, ta: tabular, f: force sensor, p: proprioception sensor, s: set, o: optical flow), and 20 prediction tasks.

Research Area	Size	Dataset	Modalities	# Samples	Prediction task
Affective Computing	S	MUSTARD (Castro et al., 2019)	$\{\ell, v, a\}$	690	sarcasm
	M	CMU-MOSI (Zadeh et al., 2016)	$\{\ell, v, a\}$	2,199	sentiment
	L	UR-FUNNY (Hasan et al., 2019)	$\{\ell, v, a\}$	16,514	humor
	L	CMU-MOSEI (Zadeh et al., 2018)	$\{\ell, v, a\}$	22,777	sentiment, emotions
Healthcare	L	MIMIC (Johnson et al., 2016)	$\{t, ta\}$	36,212	mortality, ICD-9 codes
Robotics	M	MUJOCO PUSH (Lee et al., 2020)	$\{i, f, p\}$	37,990	object pose
	L	VISION&TOUCH (Lee et al., 2019)	$\{i, f, p\}$	147,000	contact, robot pose
Finance	M	STOCKS-F&B	$\{t \times 18\}$	5,218	stock price, volatility
	M	STOCKS-HEALTH	$\{t \times 63\}$	5,218	stock price, volatility
	M	STOCKS-TECH	$\{t \times 100\}$	5,218	stock price, volatility
HCI	S	ENRICO (Leiva et al., 2020)	$\{i, s\}$	1,460	design interface
Multimedia	M	MM-IMDB (Arevalo et al., 2017)	$\{\ell, i\}$	25,959	movie genre
	M	AV-MNIST (Vielzeuf et al., 2018)	$\{i, a\}$	70,000	digit
	L	KINETICS400 (Kay et al., 2017)	$\{v, a, o\}$	306,245	human action

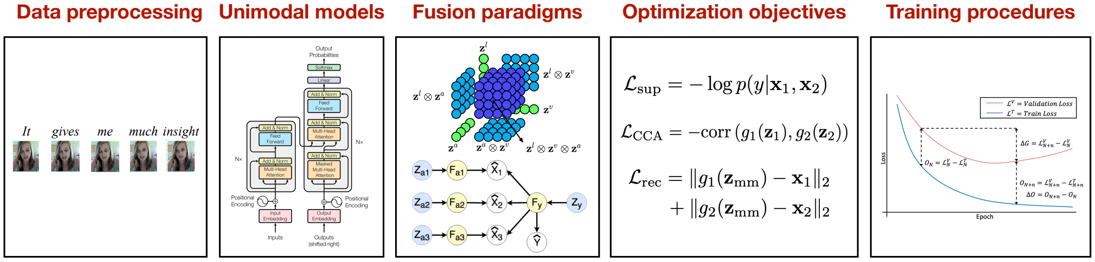
Data preprocessing
It gives me much insight
Unimodal models
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
N×
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Fusion paradigms
$\mathbf{z}^l$
$\mathbf{z}^l \otimes \mathbf{z}^v$
$\mathbf{z}^l \otimes \mathbf{z}^a$
$\mathbf{z}^v$
$\mathbf{z}^a$
$\mathbf{z}^a \otimes \mathbf{z}^v$
$\mathbf{z}^l \otimes \mathbf{z}^v \otimes \mathbf{z}^a$
Z_{a1} F_{a1} $\hat{X}_1$
Z_{a2} F_{a2} $\hat{X}_2$ F_y Z_y
Z_{a3} F_{a3} $\hat{X}_3$ $\hat{Y}$
Optimization objectives
$\mathcal{L}_{\text{sup}} = -\log p(y|\mathbf{x}_1, \mathbf{x}_2)$
$\mathcal{L}_{\text{CCA}} = -\text{corr}(g_1(\mathbf{z}_1), g_2(\mathbf{z}_2))$
$\mathcal{L}_{\text{rec}} = \|g_1(\mathbf{z}_{\text{mm}}) - \mathbf{x}_1\|_2 + \|g_2(\mathbf{z}_{\text{mm}}) - \mathbf{x}_2\|_2$
Training procedures
$\mathcal{L}^V = Validation\ Loss$
$\mathcal{L}^T = Train\ Loss$
$\Delta G = \mathcal{L}^V_{N+n} - \mathcal{L}^V_N$
$O_N = \mathcal{L}^V_N - \mathcal{L}^T_N$
$O_{N+n} = \mathcal{L}^V_{N+n} - \mathcal{L}^T_{N+n}$
$\Delta O = O_{N+n} - O_N$
Loss
Epoch

Figure 3: MULTIZOO provides a standardized implementation of multimodal methods in a modular fashion to enable accessibility for new researchers, compositionality of approaches, and reproducibility of results.

included are designed to span a range of compute times from 1 minute to 6 hours, memory from 2GB to 12GB, models from 0.01 million to 280 million parameters, and datasets from 690 to 147,000 samples. (3) Robustness: The toolkit includes both *modality-specific imperfections* taking into account each modality's unique noise topologies (i.e., flips and crops of images, natural misspellings in text, abbreviations in spoken audio), and *multimodal imperfections* across modalities (e.g., missing modalities, or a chunk of time missing in time-series data) (Liang et al., 2019; Pham et al., 2019).

Installation, testing, and integration: Our documentation provides installation instructions in Linux, MacOS, and Windows. We also include a suite of unit tests (testing self-contained functions) and integration tests (testing multiple components from across the unimodal, fusion, and training loop modules together) with 100% coverage for self-contained functions and 88% coverage overall including integration tests. We also include instructions for continuous integration: our software is hosted on GitHub which enables version control and integration via pull requests and merges. We enabled GitHub Actions workflows, which automatically runs the test builds and is triggered every time new changes are incorporated. After making the desired changes and making sure all tests pass, users can create a pull request and the authors will merge these changes into the main branch.

Together: In Algorithm 1, we show a sample code snippet in Python that loads a dataset, defines the unimodal and multimodal architectures, optimization objective, and training procedures, before running the evaluation protocol. Our toolkit is easy to use and trains models in less than 10 lines of code. By standardizing the implementation of each module and disentangling individual modules, optimizations, and training, MULTIZOO ensures accessibility and reproducibility of its algorithms.

Table 2: MultiZoo provides a standardized implementation of the following multimodal methods spanning data processing, fusion paradigms, optimization objectives, and training procedures, which offer complementary perspectives towards tackling multimodal challenges in alignment, complementarity, and robustness.

Category	Method	Alignment	Complementarity	Robustness
Data	WordAlign (Chen et al., 2017)	✓	✗	✗
Model	EF, LF (Baltrušaitis et al., 2018)	✗	✓	✗
	TF (Zadeh et al., 2017), LRTF (Liu et al., 2018)	✗	✓	✗
	MI-Matrix, MI-Vector, MI-Scalar (Jayakumar et al., 2020)	✗	✓	✗
	NL Gate (Wang et al., 2020)	✗	✓	✗
	MulT (Tsai et al., 2019a)	✓	✓	✗
	MFAS (Pérez-Rúa et al., 2019)	✗	✓	✗
Objective	CCA (Andrew et al., 2013)	✓	✗	✗
	RefNet (Sankaran et al., 2021)	✓	✗	✗
	MFM (Tsai et al., 2019b)	✗	✓	✗
	MVAE (Wu and Goodman, 2018)	✗	✓	✗
	MCTN (Pham et al., 2019)	✗	✗	✓
Training	GradBlend (Wang et al., 2020)	✗	✓	✓
	RMFE (Gat et al., 2020)	✗	✓	✓

Algorithm 1 PyTorch code integrating MultiBench datasets and MultiZoo models.

```
from datasets.get_data import get_dataloader
from unimodals.common_models import ResNet, Transformer
from fusions.common_fusions import MultInteractions
from training_structures.gradient_blend import train, test

# load Multimodal IMDB dataset
traindata, validdata, testdata = get_dataloader('multimodal_imdb')
out_channels = 3
# define ResNet and Transformer unimodal encoders
encoders = [ResNet(in_channels=1, out_channels=3, layers=5),
            Transformer(in_channels=1, out_channels=3, layers=3)]
# define a Multiplicative Interactions fusion layer
fusion = MultInteractions([out_channels*8, out_channels*32], out_channels*32, 'matrix')
classifier = MLP(out_channels*32, 100, labels=23)
# train using Gradient Blend algorithm
model = train(encoders, fusion, classifier, traindata, validdata,
        epochs=100, optimtype=torch.optim.SGD, lr=0.01, weight_decay=0.0001)
# test
performance, complexity, robustness = test(model, testdata)
```

3. Results

MultiZoo and MultiBench enable quick experimentation of multimodal algorithms for performance while balancing complexity and robustness. They uncover several shortcomings of current models, including poor generalization to out-of-domain tasks, tradeoffs between performance and efficiency, and lack of robustness to real-world imperfections. Our resources also pave the way toward answering novel research questions in multimodal transfer learning, multi-task learning, co-learning, pre-training, and interpretability. We include these results and discussions in our full paper (Liang et al., 2021) as well as scripts to reproduce these results in MultiBench software.

4. Conclusion

In conclusion, we present MultiZoo and MultiBench, a large-scale open-source toolkit unifying previously disjoint efforts in multimodal research with a focus on ease of use, accessibility, and reproducibility, thereby enabling a deeper understanding of multimodal models. Through its unprecedented range of research areas, datasets, modalities, tasks, and evaluation metrics, our toolkit paves the way toward building more generalizable, lightweight, and robust multimodal models.

Acknowledgements

This material is based upon work partially supported by Meta, National Science Foundation awards 1722822 and 1750439, and National Institutes of Health awards R01MH125740, R01MH132225, R01MH096951 and R21MH130767. PPL is partially supported by a Facebook PhD Fellowship and a Carnegie Mellon University's Center for Machine Learning and Health Fellowship. RS is supported in part by ONR N000141812861, ONR N000142312368 and DARPA/AFRL FA87502321015. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NIH, Meta, Carnegie Mellon University's Center for Machine Learning and Health, ONR, DARPA, or AFRL, and no official endorsement should be inferred. We are extremely grateful to the editor and reviewers for their valuable feedback. Finally, we acknowledge NVIDIA's GPU support.

References

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

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *5th International conference on learning representations 2017 workshop*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, 2019.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

Bruno Dumas, Denis Lalanne, and Sharon Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. In *Human machine interaction*, pages 3–26. Springer, 2009.

C. A. Frantzidis, C. Bratsas, M. A. Klados, E. Konstantinidis, C. D. Lithari, A. B. Vivas, C. L. Papadelis, E. Kaldoudi, C. Pappas, and P. D. Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, March 2010.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, 2019.

Markus A. Hollerer, Dennis Jancsary, and Maria Grafstrom. A picture is worth a thousand words: Multimodal sensemaking of the global financial crisis. *Organization Studies*, 2018.

Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rylnK6VtDH`.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Elsa A Kirchner, Stephen H Fairclough, and Frank Kirchner. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. In *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*, pages 523–576. 2019.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

Michelle A Lee, Brent Yi, Roberto Martín-Martín, Silvio Savarese, and Jeannette Bohg. Multimodal sensor fusion with differentiable filters. *IROS*, 2020.

Luis A Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile ui designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'20 Extended Abstracts)*, 2020.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, 2019.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=ttzypy3kT7`.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

M. Naphade, J. R. Smith, J. Tesic, Shih-Fu Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 2006.

Zeljko Obrenovic and Dusan Starcevic. Modeling multimodal human-computer interaction. *Computer*, 37(9):65–72, 2004.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6966–6975, 2019.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, 2019.

Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 2010.

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

Valentin Radu, Nicholas D Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 185–188, 2016.

Sethuraman Sankaran, David Yang, and Ser-Nam Lim. Multimodal fusion refiner networks. *arXiv preprint arXiv:2104.03435*, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019a.

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

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion, 2018.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5575–5585, 2018.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, et al. Multimodal machine learning for automated icd coding. In *Machine Learning for Healthcare Conference*, pages 197–215. PMLR, 2019.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2018.