Improving Language Understanding via Privileged Multimodal Training

JULLIAN A. YAPETER* and JASON ISRAEL*, University of Southern California, USA

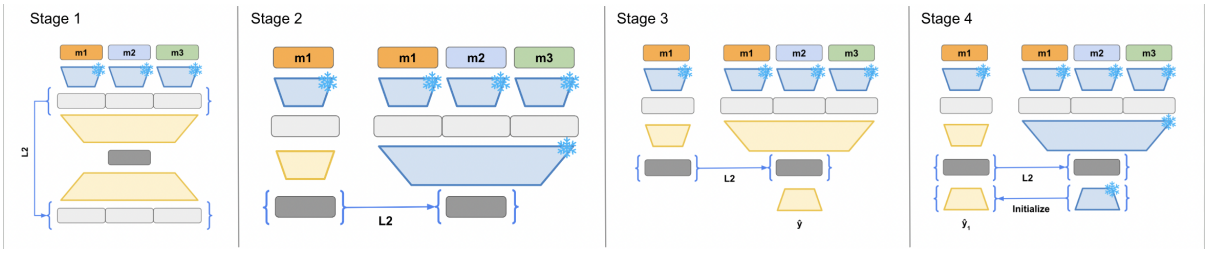

Fig. 1. An illustration of the four stages of our proposed method, from left to right. Stage 1 shows the pre-training of the multimodal network. Stage 2 shows the pre-training of the unimodal network with respect to the multimodal teacher network. Stage 3 shows the process of task-specific training, jointly optmizing the two networks. And stage 4 shows our regularized fine-tuning of the unimodal network.

ACM Reference Format:
Jullian A. Yapeter and Jason Israel. 2023. Improving Language Understanding via Privileged Multimodal Training. 1, 1 (May 2023), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Human communication is a complex and multifaceted process that involves not only linguistic cues but also various nonverbal signals, such as facial expressions, gestures, and tone of voice. Understanding and interpreting these multimodal cues is essential for effective communication, yet it remains a challenging task for both humans and machines. Privileged multimodal training is a promising approach that utilizes additional information during training that is not available during testing, aiming to improve the model's ability to generalize on communication understanding tasks. Knowledge distillation [8], on the other hand, transfers knowledge from a larger, more complex model to a smaller, simpler model, resulting in a more efficient and interpretable model.

In this project, we propose a novel approach that combines privileged multimodal training and knowledge distillation to improve human communication understanding. Specifically, we investigate how privileged information, such as speech prosody and visual data, can be incorporated into a multimodal teacher model and distilled into a unimodal student model for effective representation learning. Our goal is to achieve improved performance on unimodal downstream tasks, including emotion recognition and conversation analysis. Our findings suggest that multimodal pre-training followed by latent alignment can meaningfully improve the performance of unimodal communication models, indicating its potential for enhancing human-machine interaction and advancing our

*Both authors contributed equally to this research.

Authors' address: Jullian A. Yapeter, yapeter@usc.edu; Jason Israel, jfisrael@usc.edu, University of Southern California, Los Angeles, California, USA, 90007.

XXXX-XXXX/2023/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

understanding of human communication. This project contributes to the ongoing efforts to bridge the gap between human and machine communication and promote more effective and naturalistic interactions in various domains. The code for this paper is publicly available and can be found at https://github.com/jullian-yapeter/JudgeNet.

2 LITERATURE REVIEW

Multimodal deep learning, leveraging complementary information from paired data from different modalities, has been a long-standing topic in the field of human communication. Research in this area has given rise to different multimodal learning approaches, including Joint Multimodal Learning, Coordinated Multimodal Representation Learning, and Privileged Multimodal Learning.

Joint Multimodal Learning methods learn to leverage multimodal data to achieve improved results on downstream tasks as compared to their unimodal counterparts. Methods such as ones introduced by [13][11][10][19] utilize all available modalities at both training time and test time. This makes a strong assumption that all the data modalities used in training will be available at test time, limiting the methods' effectiveness in downstream settings where only a subset, or even just one, of the modalities are available. It has been shown by Ma *et al.* [12] that joint multimodal models are ineffective in the presence of modal-incomplete data. Our method, while learning to leverage multimodal data during training, is also explicitly optimized to work in unimodal settings. This makes our method more robust to be deployed in real-world applications where there could be missing modalities at test time.

Coordinated Multimodal Learning methods learn to embed data of different modalities into a shared latent space. This allows models to efficiently learn cross-modal concepts, embedding instances of one modality closely to instances of a different modality with similar semantic meaning. Thus, learning from data in one modality inherently supervises its learning in another. Works such as [3][14][15] make use of this concept to learn semantic representations that allow for better downstream performance in an array of multimodal tasks including cross-modal retrieval and classification. Most works, though, focus only on bimodal settings (e.g. visual-lexical, audio-lexical). In this work, we apply our method to both the bimodal and trimodal cases. Additionally, coordinated multimodal models are not designed to take joint multimodal data as input if it becomes available at test time. Our proposed method is trivially extensible to any number of modalities, and is explicitly trained to make use of joint multimodal data at test time if available.

Privileged Multimodal Learning works such as [16][17][4][18][6] are closest to our work in terms of its problem formulation. These methods make use of joint multimodal data during training, but optimize for unimodal inference. These methods often include a knowledge distillation step which uses a multimodal network as a teacher to a unimodal student network. Our work makes use of existing ideas in this line of work and propose additional techniques such as a two-step self-supervised task-agnostic pre-training process for improved initialization of the unimodal student network, as well as a novel regularization objective to guide the optimization of the unimodal task-specific head.

3 DATA

In our work, we have utilized three datasets. They are IEMOCAP, the MIT Interview Dataset, and a self-constructed dataset of TED Talks.

3.1 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [2] is a publicly available multimodal dataset that contains recordings of naturalistic dyadic interactions between actors in various emotional states. The dataset includes approximately 12 hours of audio-visual recordings of ten actors (five male and five female)

engaged in both scripted and improvised dialogue. The actors were instructed to engage in various emotional scenarios, including anger, happiness, sadness, and neutral.

The IEMOCAP dataset includes synchronized recordings of facial expressions, body movements, speech, and physiological signals. It also contains dimensional attributes such as valence, activation and dominance. It was annotated for 11 different emotional states by multiple annotators. For visual features, we are using face embedding obtained from a ResNet model which was pre-trained on ImageNet. For acoustic features, we have embeddings extracted by VGGish [7] and compressed with PCA. For lexical features, we are using BERT [5] encodings of the transcripts.

3.2 MIT Interview

The MIT Interview Dataset [9] is publicly available and contains recordings of job interviews conducted between job candidates and interviewers. The dataset was developed by researchers at MIT and includes about 10 hours of audio and video recordings of 40 mock job interviews conducted between college students and trained interviewers. The interviews cover a wide range of job positions and were designed to simulate real-world job interviews. The dataset also includes interview transcripts, obtained using a combination of ASR and manual transcription.

The MIT Interview Dataset includes annotation of various social signals, such as turn-taking, eye gaze, and facial expressions, which can provide insights into the dynamics of human communication. The dataset is multimodal in nature, containing lexical, prosodic, and visual information. The audio features were extracted using VGGish. We extracted the lexical features from the transcripts using the uncased base BERT model. Each participant was rated on a 1-7 scale by 10 Mechanical Turk annotators on a variety of traits such as excitement, engagement, eye contact, filler word usage, and recommendation for hiring. We are using the excitement score (normalized between 0-1) averaged across annotators as the target.

3.3 TED Talk

TED Talks are a series of online videos that features presentations on a wide range of topics, including science, technology, entertainment, and design. The talks are typically given by experts in their field and are designed to be engaging, informative, and inspiring. TED Talks are widely recognized as a valuable resource for learning and personal development, with millions of people around the world watching, sharing, and rating the videos. TED Talks are also an excellent resource for research into multimodal human communication. They feature a variety of speakers covering diverse topics, providing a rich source of data for studying human communication in the context of a presentation to a large audience. In addition to the speakers' chosen words, TED Talk videos also include audio and visual information that can be used for machine learning tasks.

We created a fully automated data extraction pipeline for multimodal TED Talk data and used a publicly available Kaggle dataset [1] as a starting point. The Kaggle dataset contains information about audio-video recordings of 2500 TED Talks uploaded to the official TED website up to September 21, 2017. It has links to the videos on the TED website and video metadata, the most relevant of which is a set of ratings for fourteen categories like Funny, Inspiring, Informative, Longwinded, Confusing, and Unconvincing. We synthesized video scores by categorizing the tags into positive and negative groups, and using the ratio of the two. The data pipeline scrapes the TED website for a talk's video file, and converts it to an audio file. It then extracts audio features using VGGish. Finally, it extracts lexical features from the transcripts using the base uncased BERT model.

4 METHOD

Our proposed method closely follows the teacher-student knowledge distillation paradigm, but with a few additional techniques. For each modality we first extract features in a modality-specific manner as outlined in

section 3. We have a multimodal teacher network and a unimodal student network, each of which takes as input the extracted features (in the multimodal case, it takes as input the concatenation of the unimodal features). The training process consists of four stages; a two-stage self-supervised pre-training process and a two-stage task-specific training process. A summary of the full procedure is illustrated in figure 1.

We first pre-train the multimodal network using a self-supervised autoencoder-style reconstruction objective, learning to encode multimodal features into a lower-dimensional latent-space. The goal of this first stage is to initialize the multimodal encoder with sensible weights. The loss function being minimized in the first stage is shown in equation 1, where q and p are the multimodal encoder and decoder respectively, and x is the input features. In the second stage, we pre-train the unimodal encoder to embed the unimodal data (in our case lexical or acoustic) into the same latent-space as the multimodal teacher, keeping the teacher network frozen. This allows the unimodal network to receive supervision from the paired multimodal data, helping it learn richer embeddings for the unimodal input. The loss function for the second stage is shown in equation 2, where z_i is the unimodal latent and z_f is the multimodal latent. sg is the stopgrad operator, causing the unimodal network to be influenced by the multimodal teacher and not the inverse. Note that in the first two stages, we do not require task-specific data. This means our method is able to benefit from any dataset of paired multimodal data, making it easier to gather pre-training data.

$$L = \|p(q(x)) - x\|_2 \tag{1}$$

$$L = \left\|z_i - sg(z_f)\right\|_2 \tag{2}$$

In the third stage, we unfreeze our encoders, initialize a task-specific head for the multimodal network and begin the task-specific training process. We jointly optimize the multimodal and unimodal networks by minimizing both the task-specific loss as well as a reconstruction loss unidirectionally between the unimodal encoder latent and that of the multimodal encoder. The two losses are weighted using a ratio term, α. The loss function for this third stage is shown in equation 3, where l is the task-specific loss function. This results in a trained task-specific head which can technically be used for both the unimodal and multimodal networks, since the unimodal and multimodal encoders map to the same latent space. In this work, we go one step further to optimize the unimodal case with a fourth and final stage, which we refer to as our *fine-tuning* step. We initialize a task-specific head for the unimodal network by copying the trained task-specific head of the multimodal network, and introduce a novel regularized optimization objective that both minimizes the task-specific loss and keeps the unimodal encoder close to its multimodal counterpart. This regularized objective improves the downstream performance of the unimodal network through explicit optimization, but prevents the unimodal network from straying too far from the teacher network. This discourages overfitting to the unimodal input, and improves generalization capacity. The loss function for this fourth stage is shown in equation 4. At evaluation time, we only feed unimodal input into our unimodal network.

$$L = l(\hat{y_f}, y) + \alpha \left\|z_i - sg(z_f)\right\|_2 \tag{3}$$

$$L = l(\hat{y_i}, y) + \alpha \left\|z_i - sg(z_f)\right\|_2 \tag{4}$$

To evaluate our method, we ran our full procedure with all three datasets. The downstream tasks for each of the datasets are 4-class emotion classification for IEMOCAP, binary classification (top/bottom 50% of rated talks) for TED Talks, and the regression task of candidate excitement rating prediction for the MIT Interview Dataset. For each of these tasks, we experimented with both lexical and acoustic modalities as our unimodal input, exploring the resulting performance gain on each modality. In addition, we performed an ablation study to examine the effects of our pre-training and fine-tuning stages (with and without L2 latent regularization). The performance

of our method was compared to a series of baselines, including a unimodal baseline network, a multimodal early-fusion network (oracle), and a knowledge distillation baseline [8]. Lastly, we assess our method's ability to learn from task-agnostic data, by pre-training on a different dataset than that of the downstream task.

5 RESULTS

Our experiments show that both the pre-training and fine-tuning steps helped to improve the accuracy of a unimodal model on downstream tasks. In the first ablation, our procedure performed worse than the baseline because it never explicitly optimized for the unimodal objective. With all our proposed features in place, we significantly outperform the unimodal baseline in all three downstream tasks. We also found that the L2 latent regularization was helpful in most but not all cases. We believe that the latent regularization is most beneficial when the multimodal network significantly outperforms the unimodal network; regularizing against the multimodal latent in this case better guides the unimodal network's weight search. We performed the same set of experiments twice, once with lexical unimodal input, and once with acoustic. Our ablation results are summarized in tables 1 and 2, along with the unimodal and multimodal baselines.

Dataset	Unimodal Baseline	Ours w/out pre-training w/out fine-tuning	Ours w/ pre-training w/out fine-tuning	Ours w/ pre-training w/ fine-tuning w/out regularization	Ours w/ pre-training w/ fine-tuning w/ regularization	Multimodal Baseline (Oracle)
IEMOCAP (F1)	81.2%	69.9%	70.8%	82.0%	**82.6%**	**87.6%**
TED Talk (F1)	61.6%	33.4%	32.6%	**63.2%**	53.1%	**63.2%**
MIT Interview (R^2)	0.0406	0.0522	0.0601	0.0813	**0.0912**	**0.102**

Table 1. Comparing performance of our procedure to a baseline, an oracle, and three ablations on lexical modality

Dataset	Unimodal Baseline	Ours w/out pre-training w/out fine-tuning	Ours w/ pre-training w/out fine-tuning	Ours w/ pre-training w/ fine-tuning w/out regularization	Ours w/ pre-training w/ fine-tuning w/ regularization	Multimodal Baseline (Oracle)
IEMOCAP (F1)	71.2%	66.1%	68.8%	73.0%	**75.2%**	**87.3%**
TED Talk (F1)	51.7%	51.4%	52.9%	53.2%	**54.0%**	**54.5%**
MIT Interview (R^2)	0.0450	0.0433	0.0545	**0.0789**	0.0748	**0.397**

Table 2. Comparing performance of our procedure to a baseline, an oracle, and three ablations on acoustic modality

We then evaluated our model against knowledge distillation baselines for all three tasks, and across both lexical and acoustic unimodal inputs. These knowledge distillation baselines perform distillation through a weighted optimization objective of both the student network's hard-prediction task-specific loss, and the difference between the teacher's and student's soft-predictions, as presented in [8]. The loss functions of the classification and regression knowledge distillation baselines are shown in equations 5 and 6, respectively.

$$L_{KD_{classification}} = \alpha T^2 \,\mathrm{KL}\left(\mathrm{softmax}(\frac{z_u}{T}), \mathrm{softmax}(\frac{z_m}{T})\right) + (1-\alpha)CE(z_u, y) \tag{5}$$

$$L_{KD_{regression}} = \alpha \left\| \frac{\hat{y_u}}{T}, \frac{\hat{y_m}}{T} \right\|_2 + (1-\alpha)\left\|\hat{y_u}, y\right\|_2 \tag{6}$$

Our results, summarized in tables 3 and 4, show that our method is better at leveraging the multimodal data than the knowledge distillation baselines.

Avg. over 10 runs	Knowledge Distillation Baseline	Ours
IEMOCAP (F1)	82.4%	**82.6%**
TED Talk (F1)	61.8%	**63.2%**
MIT Interview (R^2)	0.0483	**0.0912**

Table 3. Comparing performance of our procedure to a knowledge distillation baseline on lexical modality

Avg. over 10 runs	Knowledge Distillation Baseline	Ours
IEMOCAP (F1)	**65.6%**	65.2%
TED Talk (F1)	53.2%	**54.0%**
MIT Interview (R^2)	0.0460	**0.0748**

Table 4. Comparing performance of our procedure to a knowledge distillation baseline on acoustic modality

Finally, we examined the impact of task-agnostic and task-specific pretraining on the performance of our model. To obtain task-agnostic results, we conducted pretraining on either the TED or IEMOCAP datasets, and evaluated on the other. We then compared these results to the performance of our model without any pretraining, as well as with pretraining on the same dataset used for evaluation. The results in tables 5 and 6 show that pretraining helped in all cases, with task-agnostic pretraining sometimes outperforming task-specific pretraining. We theorize this is because the TED dataset contains information from a wider variety of contexts than the fairly limited IEMOCAP dataset and thus is able to find a better initialization for the model. This shows that our method is more flexible than traditional knowledge distillation, allowing for task-agnostic "distillation".

Avg. over 10 runs	Ours (w/out pre-training)	Ours (w/ task-agnostic pre-training)	Ours (w/ task-specific pre-training)
TED -> IEMOCAP (F1)	83.2%	**84.1%**	83.8%
IEMOCAP -> TED (F1)	74.5%	80.6%	**80.4%**

Table 5. Comparing performance of our procedure with task-agnostic, task-specific, and no pretraining on lexical modality

Avg. over 10 runs	Ours (w/out pre-training)	Ours (w/ task-agnostic pre-training)	Ours (w/ task-specific pre-training)
TED -> IEMOCAP (F1)	73.1%	73.7%	**76.4%**
IEMOCAP -> TED (F1)	52.6%	**53.3%**	53.1%

Table 6. Comparing performance of our procedure with task-agnostic, task-specific, and no pretraining on acoustic modality

In conclusion, we found that our method is capable of leveraging multimodal information at training-time for improved performance on downstream unimodal tasks. Our proposed pre-training and regularized fine-tuning steps both proved to be beneficial. Moreover, our results show that pre-training on a latent reconstruction objective enabled knowledge distillation at a task-agnostic level, outperforming traditional knowledge distillation methods.

6 TEAM MEMBERS' CONTRIBUTIONS

Jullian	Jake
Problem Formulation	Problem Formulation
Model architecture	TED talk & MIT data collection/processing
Implement model & ablations	Implement baselines
Implement Knowledge Distillation baseline	Data loading functionality
Develop training and testing pipelines	Integrating models into pipeline
Slide deck	Collect results
Run experiment variations	Run experiment variations
Hyperparameter tuning	Hyperparameter tuning
Write final report	Write final report

REFERENCES

[1] Rounak Bantik. 2017. *TED Talks Dataset.* Technical Report. Kaggle Datasets. url https://www.kaggle.com/datasets/rounakbanik/ted-talks .

[2] C. Lee A. Kazemzadeh E. Mower S. Kim J. Chang S. Lee C. Busso, M. Bulut and S. Narayanan. 2008. *IEMOCAP: Interactive emotional dyadic motion capture database.* Technical Report. Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359. url https://sail.usc.edu/iemocap/iemocap$_r elease.htm$.

[3] Chen Chen, Nana Hou, Yuchen Hu, Heqing Zou, Xiaofeng Qi, and Eng Siong Chng. 2022. Interactive Audio-text Representation for Automated Audio Captioning with Contrastive Learning. arXiv:2203.15526 [cs.SD]

[4] Jin Y Liu Q Heng PA. Chen C, Dou Q. 2022. Learning With Privileged Multimodal Knowledge for Unimodal Segmentation. IEEE Trans Med Imaging., 6621–632. https://doi.org/10.1109/TMI.2021.3119385

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[6] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. 2020. Learning with Privileged Information via Adversarial Discriminative Modality Distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (oct 2020), 2581–2593. https://doi.org/10.1109/tpami.2019.2929038

[7] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. arXiv:1609.09430 [cs.SD]

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML]

[9] IEEE Md. Iftekhar Tanveer Student Member IEEE Daniel Gildea Iftekhar Naim, Student Member and IEEE Mohammed Ehsan Hoque, Member. 2018. *Automated Analysis and Prediction of Job Interview Performance.* Technical Report. IEEE Transactions on Affective Computing.

[10] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2915–2921. https://doi.org/10.1145/3308558.3313552

[11] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334 [stat.ML]

[12] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are Multimodal Transformers Robust to Missing Modality? arXiv:2204.05454 [cs.CV]

[13] Asif Iqbal Middya, Baibhav Nag, and Sarbani Roy. 2022. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems* 244 (2022), 108580. https://doi.org/10.1016/j.knosys.2022.108580

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[15] Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. arXiv:1906.04402 [cs.CV]

[16] Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. 2022. Representation Learning Through Cross-Modal Conditional Teacher-Student Training For Speech Emotion Recognition. 6442–6446. https://doi.org/10.1109/ICASSP43922.2022.9747754

[17] Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. VidLanKD: Improving Language Understanding via Video-Distilled Knowledge Transfer. arXiv:2107.02681 [cs.CL]

[18] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. 2020. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1828–1838. https://doi.org/10.1145/3394486.3403234

[19] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115