A comparison of shared encoders for multimodal emotion recognition

Sai Anuroop Kesanapalli, Riya Ranjan, Aashi Goyal, Wilson Tan

CSCI 535: Multimodal Probabilistic Learning of Human Communication
Spring 2024

University of Southern California

Table of Contents

Problem Definition

- Multimodal learning aims to create models that process and relate information from multiple modalities.
- Human communication is multimodal by nature which limits the performance of unimodal models.
- A shared encoder architecture may be capable of fusing multimodal information while providing better synergy between modalities compared to architectures that use separate encoders.

Table of Contents

Background

- Buddi et al. [2] provide architectures that have one encoder tailored per modality. These are specific to voice assistants on smart-watches that utilize accelerometer readings and audio cues. We wish to use a common encoder rather than independent ones.
- Lei et al. [10] leverage the benefits of complementary information provided by different types of labels and develop three ranking models based on SVM, DNN, and GBDT.This direction is orthogonal to our approach, yet an interesting one to consider since their task is emotion recognition as well.
- Li et al. [11] propose one sensor fusion model that is designed for Radar and Lidar data, both of which are visual in nature. Moreover they employ a student-teacher framework. Despite the differences, our work draws inspiration from their sensor fusion pipeline, albeit customized for audio-visual data in our case.

Background (cont'd.)

- Yin et al. [20] propose a method where normalization parameters are exchanged between modes for implicit feature alignment. However they too employ one encoder per modality.
- Liang et al. [12] propose HighMMT, an architecture scalable with modalities. Our pipelines share structural similarities with HighMMT. Albeit we employ multiple classes of shared encoders, such as 2D CNN, 3D CNN, and Transformer, rather than devising a customized Transformer-based architecture.

Table of Contents

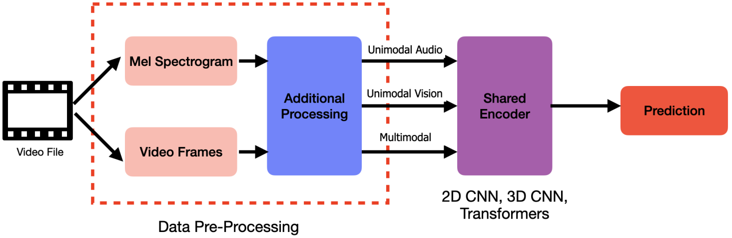
Data

- As a proof of concept, we wish to test this architecture for emotion recognition on CREMA-D dataset [3], given its simplicity and aptness for our bimodal use-case.
- Evaluated by over $2,400$ individuals, CREMA-D includes $7,442$ video clips with performances by 91 actors, providing a diverse exploration of emotional expression. 7439 for 3D experiments.
- Within the dataset, each actor presents 12 sentences, expressing 6 emotions at different intensity levels.
- Each video clip is brief, lasting less than 5 seconds.

Table of Contents

Method
Architecture

- Videos are pre-processed to generate frames containing faces, Mel spectrograms, concatenated if the pipeline is multimodal, along with additional processing depending on the architectural requirements of the encoder, then passed-on to the encoder which performs emotion recognition.
- This architectural design draws inspiration from the plug-and-play ideology, with shared encoder being the changeable component.



- For pre-processing, frames were extracted from videos and resized to 224×224 images. Middle frame was chosen to perform face-detection using a MTCNN, and the frame was then cropped to the detected face.
- For audio pre-processing, Mel spectrograms were generated using `librosa` at a sample rate of $22,050$ Hz, 2048 FFT points, hop length of 512, and 512 Mel bands. These spectrograms are then resized to 224×224 images.
- In the multimodal pipeline, these faces and spectrogram images are concatenated horizontally to form a single chunk of multimodal data, which is the passed-on to the encoder employed in the pipeline (2D CNN, ViT).

Method
3D Experiments

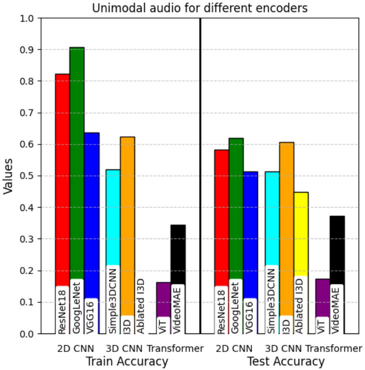
- For pre-processing, Mel spectrograms were evenly divided into chunks along the time-axis. The number of chunks they were divided into varied to match the number of frames extracted from their corresponding video. This was done to temporally align frames with spectrogram chunks. Mel spectrogram chunks were then resized to 224×224 images. These were used as the 3D unimodal vision and audio data.
- Now that frames and spectrogram chunks are temporally aligned, they were horizontally concatenated together to form the 3D multimodal data. After concatenation, each combined frame and spectrogram chunk formed an image of size 448×224.
- For the 3D transformer multimodal data, video frames were further resized to 208×224 and spectrogram chunks were resized to 16×224 before concatenation. After concatenation, they formed images of size 224×224.

Table of Contents

Results

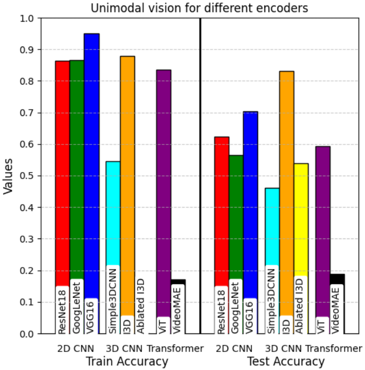
Unimodal audio

Type	Encoder	Train Acc.	Test Acc.
2D CNN	ResNet18	0.8216	0.5825
	GoogLeNet	0.907	**0.619**
	VGG16	0.6366	0.5120
3D CNN	Simple3D CNN	0.519	0.514
	I3D	0.623	0.605
	Ablated I3D	-	0.448
Trans-former	ViT	0.1634	0.1738
	VideoMAE	0.344	0.372



Results
Unimodal vision

Type	Encoder	Train Acc.	Test Acc.
2D CNN	ResNet18	0.8634	0.6225
	GoogLeNet	0.866	0.566
	VGG16	0.9495	0.7040
3D CNN	Simple3D CNN	0.546	0.462
	I3D	0.878	**0.831**
	Ablated I3D	-	0.540
Transformer	ViT	0.8361	0.5934
	VideoMAE	0.170	0.188



Results
Multimodal

Type	Encoder	Train Acc.	Test Acc.
2D CNN	ResNet18	0.8854	0.6350
	GoogLeNet	0.925	0.661
	VGG16	0.4329	0.4716
3D CNN	Simple3D CNN	0.647	0.573
	I3D	0.923	**0.806**
Trans-former	ViT	0.6811	0.5598
	VideoMAE	0.334	0.366

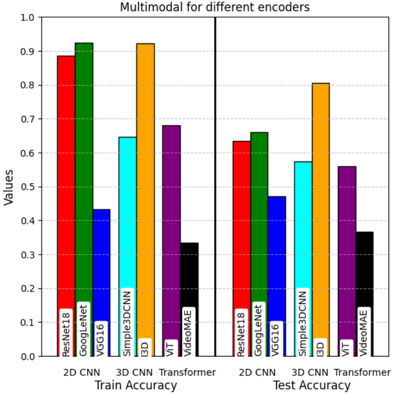


Table of Contents

- **CNNs:**
  1. The classic CNNs – ResNet18, VGG16, and GoogLeNet perform decently on the test split, with GoogLeNet outperforming others in unimodal audio and multimodal scenarios, and VGG16 doing the best in case of unimodal vision.
  2. Of the three modalities, audio accuracies are considerably lower than the rest. This is probably due to two reasons – much information regarding emotion of the speaker is not contained in the audio when compared to vision, and Mel spectrogram conversion may be leading to loss of some information.
- **ViT:**
  1. Contrary to our initial guess, ViT does not always perform better than 2D CNNs. An explanation for this lies in the observation that ViTs are known to outperform CNNs, but only when trained on large datasets (14-300M images).
  2. In the audio modality, it performs much worse when compared to vision and multimodal scenarios, owing to the patching scheme which is not suitable for Mel spectrograms.

Discussion
3D

- **Video Transformers:**
  1. Although video transformers show good results on other datasets like Kinetics, they struggle with spatial redundancy which Kinetics mitigates with diverse actions and environments.
  2. Furthermore, joint-space attention used in VideoMAE scales quadtratically with respect to both image size and number of frames. Adding a small 3D CNN model may help mitigate both the issues.
- **3D CNNs:** Simple3D CNN is a tiny model (3262 parameters) but performs decent already.
- **Converting Audio to 3D:**
  1. Might be a waste of parameters, but it does help with multimodal interaction.
  2. Also worked well for small 3D CNN models which gave better results than some large 2D CNN models.
  3. Tuning hyperparameters would improve results.

Discussion

2D vs 3D

- **I3D** outperforms all encoders across all modalities, except for GoogLeNet in the case of unimodal audio.
- It is not the norm that 3D encoders work better than their 2D counterparts.
- In scenarios where these models are to be deployed on the edge, 2D encoders have an upper-hand due to their faster data pre-processing and training times. However where compute is not a constraint, and for mission-critical applications with low tolerance for misclassifications, 3D encoders are an ideal choice.

Table of Contents

Summary

- Developed unimodal audio and vision, and multimodal emotion recognition pipelines.
- Employed various classes of shared encoders –
  - 2D CNNs: ResNet18 ($\sim$ 11.7M), GoogLeNet ($\sim$ 7M), and VGG16 ($\sim$ 138M)
  - 3D CNNs: Simple3D CNN (3262), and I3D ($\sim$ 12.3M)
  - Transformers: ViT ($\sim$ 16.4M) and VideoMAE
- Tested our pipelines on a full-scale version of CREMA-D dataset that contains 7442 (7439) videos of actors expressing 6 kinds of emotions in various intensities.
- Presented a principled comparison of the performance of different pipelines and encoders, identified the achievements and shortcomings of these architectures, and discussed the implications along with the possibilities for future work.

Table of Contents

Member Contributions

- Anuroop
  1. Responsible for implementing unimodal audio and vision, multimodal pipelines with 2D CNN - ResNet18
  2. Midterm presentation deck and report
  3. Responsible for implementing unimodal audio and vision, multimodal pipelines with ViT, and for running the corresponding full-scale experiments for unimodal audio and multimodal pipelines
  4. Responsible for 2D data pre-processing for full-scale experiments
  5. Midterm and Final presentation deck and report – including barplots, analysis, and discussion on 2D experiments
- Riya
  1. Unimodal audio and vision pipelines with 2D CNN - GoogLeNet
  2. Multimodal pipeline with 2D CNN - GoogLeNet
  3. Fine tune GoogLeNet model on batch size and learning rates, with best fit results, analysis and discussions
  4. Testing out ViT fullscale experiments with different combinations of batch size, heads, blocks, dropout rates and analysis.
  5. Midterm and Final report

Member Contributions (cont'd.)

- Aashi
  1. Unimodal audio and vision pipelines with 2D CNN - VGG16
  2. Mulitomodal pipeline with 2D CNN - VGG16
  3. Fine-Tuning of VGG16 model, with best fit results, analysis and discussions
  4. Testing out ViT fullscale experiments with different combinations of batch size and learning rates and analysis.
  5. Midterm and Final report
- Wilson
  1. Solely responsible for all of 3D parts including: 3D data preprocessing, 3D encoder training/testing (unimodal and multimodal), 3D analysis and discussion, etc.
  2. Midterm and Final report

Table of Contents

Future Work

- ViT architecture can be further improved and trained on a much bigger dataset to match the current state-of-the-art performance.
- Patching of audio modality information encoded as Mel spectrograms is not really an ideal choice. A better thing to do is to replicate these spectrograms across the patches and concatenate these replicated spectrograms with the patched video frames.
- In the 3D pipeline, adding a small 3D CNN may help mitigate spatial redundancy in videos and also address joint-space attention used in VideoMAE that scales quadtratically with respect to both image size and number of frames.
- Each experiment can be run multiple times and the averaged metrics of these set of experiments along with error bars can be reported, as a better practice.

Table of Contents

References I

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. *Is Space-Time Attention All You Need for Video Understanding?* 2021. arXiv: 2102.05095 [cs.CV].

[2] Sai Srujana Buddi et al. *Efficient Multimodal Neural Networks for Trigger-less Voice Assistants*. 2023. arXiv: 2305.12063 [cs.LG].

[3] Houwei Cao et al. "Crema-d: Crowd-sourced emotional multimodal actors dataset". In: *IEEE transactions on affective computing* 5.4 (2014), pp. 377–390.

[4] Joao Carreira and Andrew Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2018. arXiv: 1705.07750 [cs.CV].

[5] Eric Dodds et al. *Modality-Agnostic Attention Fusion for visual search with text feedback*. 2020. arXiv: 2007.00145 [cs.CV].

References II

[6] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[7] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[8] Bin Jiang et al. "Multi-task Cascaded Convolutional Neural Networks for Real-Time Dynamic Face Recognition Method". In: *Communications, Signal Processing, and Systems*. Ed. by Qilian Liang et al. Singapore: Springer Singapore, 2020, pp. 59–66.

[9] Will Kay et al. *The Kinetics Human Action Video Dataset*. 2017. arXiv: 1705.06950 [cs.CV].

References III

[10] Yuanyuan Lei and Houwei Cao. "Audio-Visual Emotion Recognition With Preference Learning Based on Intended and Multi-Modal Perceived Labels". In: *IEEE Transactions on Affective Computing* 14.4 (2023), pp. 2954–2969. DOI: 10.1109/TAFFC.2023.3234777.

[11] Yu-Jhe Li et al. "Modality-agnostic learning for radar-lidar fusion in vehicle detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 918–927.

[12] Paul Pu Liang et al. *High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning*. 2023. arXiv: 2203.01311 [cs.LG].

[13] Arsha Nagrani et al. "Attention bottlenecks for multimodal fusion". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14200–14213.

References IV

[14] Javier Selva et al. "Video Transformers: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), pp. 1–20. ISSN: 1939-3539. DOI: 10.1109/tpami.2023.3243465. URL: http://dx.doi.org/10.1109/TPAMI.2023.3243465.

[15] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[16] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[17] Zhan Tong et al. *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. 2022. arXiv: 2203.12602 [cs.CV].

References V

[18] Zhan Tong et al. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 10078–10093. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

[19] 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*. 2020, pp. 12695–12705.

[20] Yufeng Yin et al. "X-Norm: Exchanging Normalization Parameters for Bimodal Fusion". In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. 2022, pp. 605–614.

References VI

[21] Amir Zadeh et al. "Multi-attention recurrent network for human communication comprehension". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Thank you!