A multimodal architecture with shared encoder that uses spectrograms for audio

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

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

Problem Definition

- Multimodal learning aims to create models that process and relate information from multiple modalities.

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.

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

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

Background

- Buddi et al. [1] 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.

Background

- Buddi et al. [1] 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.

Background

- Buddi et al. [1] 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. [5] leverage the benefits of complementary information provided by different types of labels and develop three ranking models based on SVM, DNN, and GBDT.

Background

- Buddi et al. [1] 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. [5] 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.

Background

- Buddi et al. [1] 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. [5] 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. [6] 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.

Background

- Buddi et al. [1] 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. [5] 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. [6] 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

- Buddi et al. [1] 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. [5] 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. [6] 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.
- Yin et al. [11] propose a method where normalization parameters are exchanged between modes for implicit feature alignment.

Background

- Buddi et al. [1] 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. [5] 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. [6] 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.
- Yin et al. [11] propose a method where normalization parameters are exchanged between modes for implicit feature alignment. However they too employ one encoder per modality.

Table of Contents

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

Data

- As a proof of concept, we wish to test this architecture for emotion recognition on CREMA-D dataset [2], 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.
- 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

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

Method

We are working on a novel audio-visual learning paradigm where audio data is represented as spectrograms, in order for the embeddings to be used with an encoder that is shared between audio and video data.

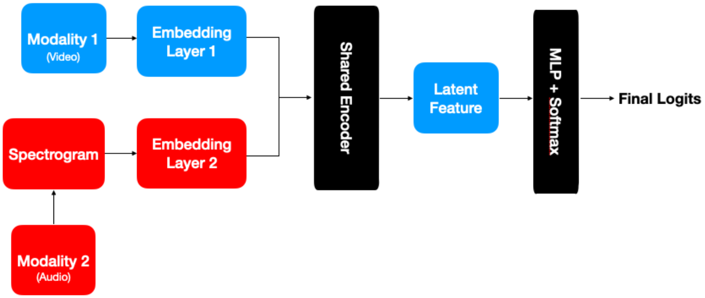

Figure: Multimodal Pipeline

Table of Contents

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

Results so far

2D-CNN — Audio / Video — Unimodal

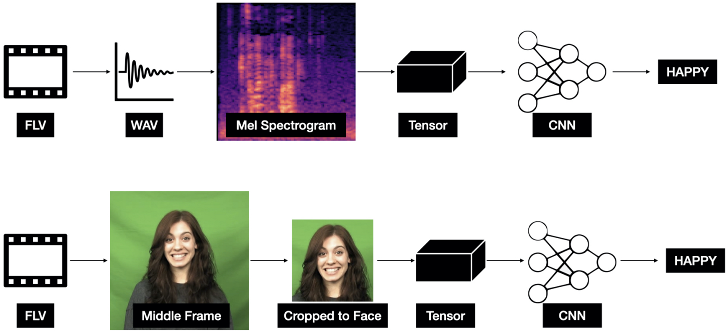

Figure: Unimodal audio and video pipelines with 2D CNN

Results so far

2D-CNN — Audio / Video — Unimodal

Type	2D CNN	Train Loss	Train Acc.	Test Loss	Test Acc.
Baseline	ResNet18	-	-	1.0953	0.3415
FT	ResNet18	0.5624	0.9895	0.6576	**0.8902**
Baseline	GoogLeNet	-	-	1.0993	0.2805
FT	GoogLeNet	0.6148	0.9319	0.8268	0.6829
Baseline	VGG16	-	-	1.1027	0.3537
FT	VGG16	0.6657	0.8848	0.7831	0.7683

Table: Unimodal Audio with 2D CNN

Results so far

2D-CNN — Audio / Video — Unimodal

Type	2D CNN	Train Loss	Train Acc.	Test Loss	Test Acc.
Baseline	ResNet18	-	-	1.0967	0.3659
FT	ResNet18	0.5809	0.9686	0.7736	**0.7805**
Baseline	GoogLeNet	-	-	1.0987	0.3171
FT	GoogLeNet	0.6992	0.8429	0.9076	0.6463
Baseline	VGG16	-	-	1.0919	0.3902
FT	VGG16	0.5927	0.9579	0.8393	0.7073

Table: Unimodal Video with 2D CNN

Results so far

2D-CNN — Audio / Video — Unimodal — Crossed

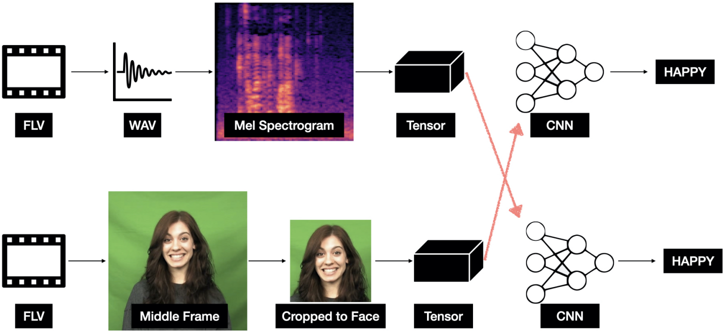

Figure: Unimodal audio and video pipelines with 2D CNN crossed

Results so far

2D-CNN — Audio / Video — Unimodal — Crossed

Type	2D CNN	Train Loss	Train Acc.	Test Loss	Test Acc.
Audio	ResNet18 (Video)	-	-	1.2287	0.3171
Video	ResNet18 (Audio)	-	-	1.2611	0.2561
Audio	GoogLeNet (Video)	-	-	1.2038	0.3293
Video	GoogLeNet (Audio)	-	-	1.1606	0.3902
Audio	VGG16 (Video)	-	-	1.1726	0.3415
Video	VGG16 (Audio)	-	-	1.2138	0.3049

Table: Unimodal crossed with 2D CNN

Results so far

2D-CNN — Audio / Video — Multimodal

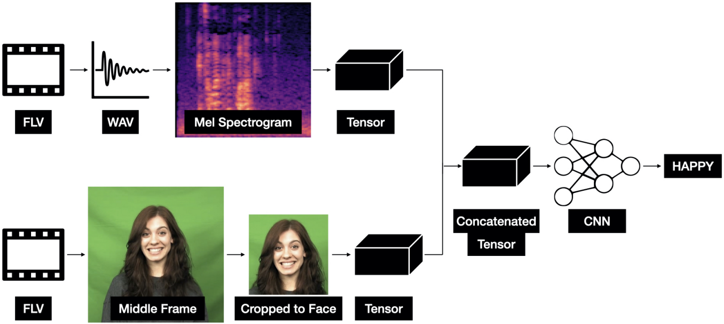

Figure: Multimodal audio and video pipelines with 2D CNN

Results so far

2D-CNN — Audio / Video — Multimodal

Type	2D CNN	Train Loss	Train Acc.	Test Loss	Test Acc.
Baseline	ResNet18	-	-	1.0839	0.3780
FT	ResNet18	0.5515	1.0000	0.7124	**0.8415**
Baseline	GoogLeNet	-	-	1.0987	0.3171
FT	GoogLeNet	0.6793	0.8639	0.8967	0.6585
Baseline	VGG16	-	-	1.1029	0.2683
FT	VGG16	0.5942	0.9579	0.9061	0.6220

Table: Multimodal with 2D CNN

Results so far

3D-CNN — Video — Unimodal

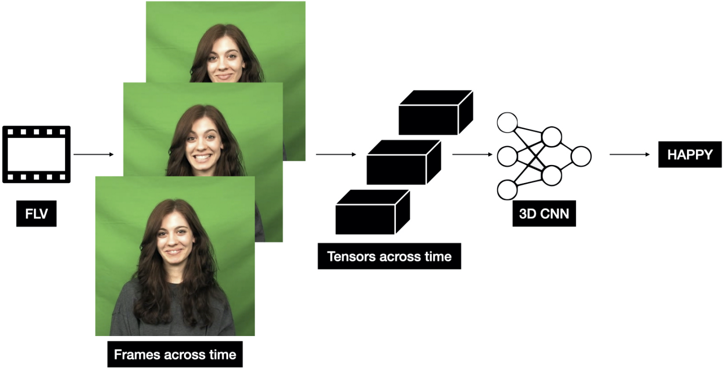

Figure: Unimodal video pipeline with 3D CNN

Results so far

3D-CNN — Video — Unimodal

Type	3D CNN	Train Loss	Train Acc.	Test Loss	Test Acc.
FT	Simple3D CNN	0.3320	0.8490	0.4763	**0.8214**
FT	I3D	0.6406	0.7061	0.8172	0.7143

Table: Unimodal video with 3D CNN

Results so far

Training setup

Parameter	ResNet18, GoogLeNet, VGG16	Simple3D CNN, I3D
# samples (train + test)	273 (191 + 82)	273 (245 + 28)
Batch size	32	8, 2
Learning rate	0.001 (0.0001 for VGG16)	0.001
Optimizer	Adam	Adam
Loss	Cross Entropy	Cross Entropy
# train epochs	50	23, 32
GPU	T4 (via Colab)	T4, T4

Table: Training setup for unimodal and multimodal pipelines with 2D CNN, and 3D CNN

Table of Contents

1. Problem Definition
2. Background
3. Data
4. Method
5. Results so far
6. Summary and member contributions
7. References

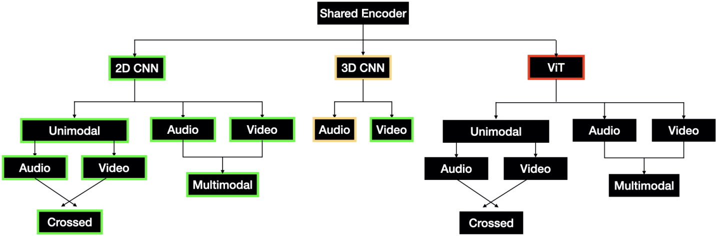

Figure: Project Status

Summary and member contributions

Summary

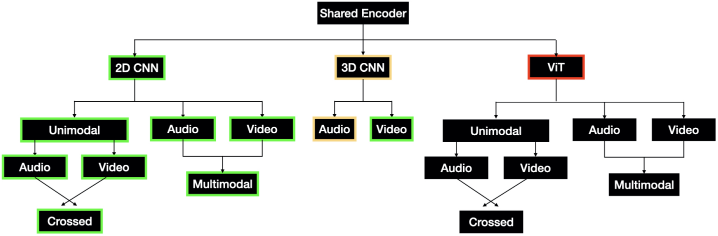

Figure: Project Status

Todo:

- Unimodal audio pipeline with 3D CNN
- Unimodal and multimodal pipelines with ViT
- Improve the existing CNN-based pipelines for better accuracy
- Scaling to full CREMA-D dataset

Summary and member contributions

Summary

- Code-base hosted on GitHub (private) repository - https://github.com/ksanu1998/multimodal_course_project.
- Our experiments are available as `.ipynb` notebooks and `.py` scripts accompanied with README files and can be reproduced.
- Please contact any of the team members for access and information.

Summary and member contributions

Member Contributions - All team members are actively involved in and contributing towards the project

- Anuroop
  1. Unimodal audio and video pipelines with 2D CNN - ResNet18
  2. Multimodal pipeline with 2D CNN - ResNet18
  3. Midterm presentation deck and report
- Riya
  1. Unimodal audio and video pipelines with 2D CNN - GoogLeNet
  2. Multimodal pipeline with 2D CNN - GoogLeNet
  3. Fine tune GoogLeNet model
- Aashi
  1. Unimodal audio and video pipelines with 2D CNN - VGG16
  2. Mulitomodal pipeline with 2D CNN - VGG16
  3. Trying out different combinations of neural networks and optimizers
- Wilson
  1. Unimodal video pipeline, 3D CNN - Simple3D CNN, 2Plus1D ResNet, I3D
  2. Still experimenting with I3D and ResNet

Table of Contents

1 Problem Definition

2 Background

3 Data

4 Method

5 Results so far

6 Summary and member contributions

7 References

References I

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

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

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

[4] 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.

References II

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] 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.

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

Thank you!