A comparison of shared encoders for multimodal emotion recognition

Sai Anuroop Kesanapalli, Riya Ranjan, Aashi Goyal, Wilson Tan
[kesanapa, riyaranj, aashiarv, wtan1167] @usc.edu
University of Southern California

1 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. In this work, a shared encoder architecture that is capable of fusing multimodal information while providing better synergy between modalities is compared to architectures that use separate encoders.

To this end, we developed unimodal audio, unimodal video, and a multimodal pipeline that builds on the former. We employ various classes of shared encoders such as 2D CNNs comprising ResNet18 [He et al., 2016], GoogLeNet [Szegedy et al., 2015], and VGG16 [Simonyan and Zisserman, 2014]; 3D CNNs comprising Simple3D CNN, and I3D [Carreira and Zisserman, 2017]; and 2D Vision Transformer (ViT) [Dosovitskiy et al., 2021] and 3D Vision Transformer (VideoMAE) [Tong et al., 2022a]. We test our pipelines on the task of emotion recognition on full-scale version of CREMA-D dataset [Cao et al., 2014] that contains 7442 videos of actors expressing 6 kinds of emotions in various intensities.

We present a principled comparison of the performance of different pipelines and encoders, identify the achievements and shortcomings of these architectures, and discuss the implications along with the possibilities for future work.

2 Literature Review

[Buddi et al., 2023] 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 and Cao, 2023] 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., 2022] propose one sensor fusion model that is designed for Radar and Lidar data, both of which are vision 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-vision data in our case. [Yin et al., 2022] propose a method where normalization parameters are exchanged between modes for implicit feature alignment. However they too employ one encoder per modality. Previous works have also leveraged attention mechanisms for fusion. [Dodds et al., 2020] present a simple modality-agnostic model by using self and cross attention on images and text to learn a common embedding space. Using transformer architectures which utilizes attention mechanisms may also be beneficial for our audio-vision task. [Liang et al., 2023] 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.

3 Data Description

We utilize the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) [Cao et al., 2014] for our work, offering a rich multimodal experience, integrating audio and video for enhanced emotion analysis. Evaluated by over 2400 individuals, CREMA-D includes 7442 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. Im-

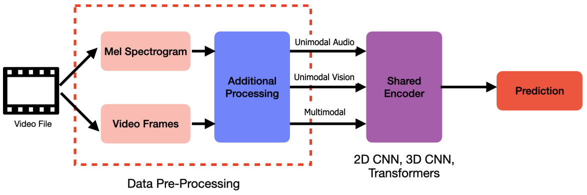


Figure 1: Architecture

portantly, the dataset includes the number of ratings for each emotion, offering valuable insights into the perceived emotional content of the performances. We identified 3 videos with recording issues which gives a total of 7439 good videos.

4 Method

Our architecture is visualized in Figure 1. Videos are pre-processed (Section 5) to generate frames and mel spectrograms. Frames and mel spectrograms are concatenated if the pipeline is multimodal, followed with additional processing depending on the architectural requirements of the encoder. It is then passed-on to the encoder which performs emotion recognition. The type of concatenation, number of channels used in the processed images, additional processing such as generation of patches for ViT and temporal alignment for 3D CNN and VideoMAE, vary depending upon the encoder employed. This architectural design draws inspiration from the plug-and-play ideology, with shared encoder being the changeable component.

5 Experiments & Results

5.1 2D Encoders

We first take a look at 2D CNNs and Vision Transformer (ViT) as shared encoders:

5.1.1 2D CNNs

ResNet18 [He et al., 2016], VGG16 [Simonyan and Zisserman, 2014], and GoogLeNet [Szegedy et al., 2015] are employed in this class of encoders. These CNNs were chosen with regards to their number of learnable parameters – GoogLeNet ($\sim$ 7M), ResNet18 ($\sim$ 11.7M), VGG16 ($\sim$ 138M), which provide us with a wide spectrum. This choice also reflects the need for consideration into deployment of this class of architectures on edge-devices in the Internet of Things (IoT) era.

5.1.2 ViT

First, a simple ViT [Dosovitskiy et al., 2021] trained from scratch on CREMA-D is employed in this experiment. This transformer accepts 1 input channel, has a patch dimension of 16, dimensionality of token embeddings is 768, has $6/8$ transformer blocks, $4/8$ attention heads, dimensionality of linear layer is 1024, and includes an additional classification token. This transformer has $\sim$ 16.4M learnable parameters (with 6 transformer blocks and 4 attention heads). We then experimented with a ViT from HuggingFace [Wu et al., 2020] pretrained on ImageNet-21k [Deng et al., 2009] at a resolution of 224×224 pixels, referred to as PT ViT hereafter.

5.2 2D Experiments

For video pre-processing, frames were extracted from videos and resized to 224×224 (Width $\times$ Height) images. Of these, the middle frame was chosen to perform face-detection using a Multi-Task Cascaded Convolutional Neural Network (MTCNN) [Jiang et al., 2020], 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 22050 Hz, 2048 FFT points, hop length of 512, and 128 mel bands. These spectrograms are then resized to 224×224 images. For ResNet18 and ViT in the unimodal and multimodal pipelines, images (both faces and spectrograms) are converted to grayscale whereas for the rest of the 2D CNNs and PT ViT pipelines, they remain multi-channeled (RGB). 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) which do further processing as per their architectural requirements. Experiments have been conducted on a full-scale version of the CREMA-D dataset with a $70/30$ train-test split, where 10 samples have been discarded as no face was detected in those, and 3 more have been discarded as they contained recording errors. This corresponds to a total of 5200 train samples and 2229 test samples. Training epochs were set to 50.

5.3 2D Results

5.3.1 Unimodal Audio

Results are described in Table 1. We observe that all the 2D CNNs – ResNet18, GoogLeNet, and

VGG16, perform similarly in comparison to each other, and better than 3D CNNs and Transformer-based encoders in general. PT ViT is observed to perform the best among this class of encoders. This is evident from the barplot in Figure 2a. We attribute this performance to the quality of pre-trained weights of these 2D encoders atop which we fine-tune using mel spectrograms. We believe ViT trained from scratch performs worse when compared to the other 2D encoders since only 5200 samples were used for training as mentioned earlier.

5.3.2 Unimodal Vision

Results are described in Table 2. We observe that ResNet18, GoogLeNet, VGG16, and PT ViT perform better in comparison to ViT. Here too PT ViT outperforms 2D CNNs by a narrow margin. This can be seen from the barplot in Figure 2b. We attribute the poor performance of ViT to the same reason mentioned in the unimodal audio scenario.

5.3.3 Multimodal

Results are described in Table 3. We observe here that PT ViT outperforms other 2D encoders, as can be seen from the barplot in Figure 2c. Among 2D CNNs, ResNet18 performs the best. Moreover, when compared to unimodal vision, test accuracies have improved slightly for ResNet18 in the multimodal scenario. A similar trend is seen for ViT and PT ViT. Interestingly, ViT in multimodal scenario is performing as good as the 2D CNNs, in contrast to the unimodal audio and vision.

5.4 2D Discussion

2D CNNs: The classic CNNs – ResNet18 [He et al., 2016], VGG16 [Simonyan and Zisserman, 2014], and GoogLeNet [Szegedy et al., 2015] perform decently on the test split, with all the 2D CNNs having similar performance in unimodal audio and vision scenarios. Train accuracies of these encoders are higher than their test accuracies. These encoders are massively overparameterized and fine-tuned on a considerably small dataset. However, this difference is not profound in case of unimodal vision and multimodal regimes. Moreover it is observed that the difference in train and test accuracies varies considerably across encoders. Of the three modalities, audio accuracies are 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: The unimodal audio and vision performance of ViT is worse when compared to multimodal scenario. Contrary to our initial assumption, ViT trained from scratch on CREMA-D 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), as mentioned in [Dosovitskiy et al., 2021]. In our case, the entire dataset consists of 7229 video clips (excluding errors), and correspondingly those many video frames as per our pre-processing scheme. For training, as mentioned earlier in this section, this number is 5200. Despite this stark difference in the number of samples required to train in order to outperform 2D CNNs, ViT does perform comparably to 2D CNNs in the multimodal case, owing to its superior architecture involving attention mechanisms – this demonstrates the benefit of using a multimodal pipeline. PT ViT, which is trained on ImageNet-21k consisting of 14M images, and fine-tuned on CREMA-D, is observed to have performed best among 2D encoders. This corroborates the aforementioned claim made by [Dosovitskiy et al., 2021], and points to the efficacy of using pre-trained models fine-tuned on target dataset.

General Discussion: Considering training time as a proxy, ResNet18 takes about 8.33 minutes to train on multimodal data for 50 epochs on an NVIDIA Tesla P100 GPU. We explored different sets of hyperparameters (Table 4) for each class of 2D encoders and modalities, such as batch size, learning rate, dropout rate (in case of ViT), and reported the values obtained that have converged at the last epoch for these encoders, except for PT ViT where we reported the best values. We fixed training epochs to 50 (30 for PT ViT), a convenient choice with regards to execution time, and ones that also correspond to a point beyond which the test accuracy does not improve further.

5.5 3D Encoders

We also look at 3D CNNs and Video Transformers as shared encoders:

Simple3D CNN: A simple CNN that uses 6 3D convolutional layers followed by a final classifcation layer. The main purpose of this model is to serve as a baseline for 3D performance.

I3D: [Carreira and Zisserman, 2017] Uses a series of inception modules, where each inception module is made up of several 3D convolutional layers. For classification, it uses average pooling over spatial and temporal dimensions to make a prediction. This model is a one-stream RGB version pretrained on ImageNet 1K.

VideoMAE: [Tong et al., 2022b] A masked autoencoder (MAE) that extends to videos by using the vanilla ViT as a backbone. It does this by masking random 3D patches in videos as opposed to 2D patches found in 2D MAEs. This model was pretrained on the Kinetics400 dataset [Kay et al., 2017].

5.6 3D Experiments

Due to recording errors in the dataset, 3 videos were removed resulting in a total of 7439 total videos. The 3D experiments use the full 7439 videos and a randomly selected 80/20 train-test split which gives a total of 5951 training and 1488 testing samples.

For video pre-processing, frames were extracted from videos at 24 frames per second and resized to 224 × 224 (Width × Height). For audio pre-processing, mel spectrograms were created using audio files then converted to 3D. To convert to 3D, the 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. Frames and spectrogram chunks retained RGB color channels. Now that frames and spectrogram chunks are temporally aligned, they are concatenated together to form the 3D multimodal data.

However, there are still two issues. Firstly, the implementation used for VideoMAE does not accept rectangular data which is an issue when frames and spectrogram chunks are concatenated as they are. This is easily resolved by further resizing video frames to 208 × 224 and spectrogram chunks to 16 × 224 before concatenation. Secondly, an issue of varying number of frames/chunks per array. This is handled differently depending on architectures:

- **3D CNNs:** Padding was used to resolve the uneven video length issue by adding blank images until all arrays had the same number of frames/chunks as the longest array (135).
- **VideoMAE:** Instead of padding, 32 frames/chunks were taken evenly spread across the number of frames/chunks to maintain good temporal fidelity while substantially lowering memory usage.

Following an NCWH format, the final data dimensions look as follows:

CNNs Unimodal: $(135, 3, 244, 244)$
CNNs Multimodal: $(135, 3, 244, 488)$
VideoMAE Unimodal: $(32, 3, 244, 244)$
VideoMAE Multimodal: $(32, 3, 244, 244)$

Early termination is used during training. Early termination was determined by monitoring model performance every epoch for signs of convergence or overfitting. The model with the best test accuracy was reported.

5.7 3D Results

In addition to comparing 3D models against each other, we will also compare them to human performance. [Cao et al., 2014] provides human performance on the CREMA-D dataset for audio-only, vision-only, and audio-vision emotion classification at 40.9%, 58.2% and 63.6% respectively.

5.7.1 Unimodal Audio

Results are shown in Table 1. Both 3D CNN models outperformed humans on unimodal audio emotion classification. Out of the two models, I3D performed better by $\sim 9\%$. However, Simple3D is by far the smaller model with only 3262 parameters compared to I3D with 12.3M parameters.

VideoMAE performs worse than Simple3D and humans, but better than random guessing. This is likely due to spatial redundancy which will be elaborated in the 3D discussion section.

5.7.2 Unimodal Vision

Results are shown in Table 2. Although Simple3D was unable to outperform humans in unimodal vision, I3D significantly does. It is likely that the model was able to transfer learn from ImageNet pretraining to boost unimodal vision performance.

VideoMAE still suffers from spatial redundancy but even more so in the vision domain. More on this in the 3D discussion section. It has similar performance to random guessing which suggests that VideoMAE was unable to learn anything useful.

5.7.3 Multimodal

Results are shown in Table 3. I3D still performs well and better than humans, but it does not do better than its unimodal vision variant, which could mean that classification is largely skewed by vision. One simple way to check is to look at model performance with modality ablation. This is done by masking a modality in the multimodal data. These results are represented as Ablated I3D. Results show a significant drop in performance when either vision or audio is removed (Tables 1 and 2 respectively). This suggests that both modalities are important for I3D multimodal classification and there is no overly dominant modality.

Simple3D again is unable to outperform humans. However, it was able use multimodal interaction to get a higher test accuracy compared to any of its unimodal versions. Judging from these results, it is unlikely that there is an overly dominant modality and that the model is truly learning multimodal interactions. However, we still verify with modality ablation. Comparing the multimodal Simple3D results with the Ablated Simple3D results shows that the model is learning interactions between vision and audio.

As for VideoMAE, because the multimodal results are similar to the unimodal audio, and because it was randomly guessing on unimodal vision, it is likely that audio is the completely dominant modality. In light of this, we chose to skip modality ablation on VideoMAE.

5.7.4 3D Discussion

Video Transformers: Although video transformers have good results on other datasets like Kinetics, they struggle with spatial redundancy [Selva et al., 2023] which Kinetics mitigates with diverse actions and environments [Kay et al., 2017]. Spatial redundancy is inherently an issue with videos, but it is especially challenging for facial emotion recognition where facial action units may persist for the duration of the emotional state. In the frequency domain, this is mitigated slightly, but still not enough to give good predictions. Furthermore, joint-space attention used in VideoMAE scales quadtratically with respect to both image size and number of frames [Bertasius et al., 2021]. Adding a small 3D CNN model may help mitigate both issues.

A 3D CNN can be used to recognize important temporal and spatial features. Not only would this shrink the temporal and spatial dimensions through convolution and pooling, this may also mitigate some effects of spatial redundancy. Although adding a 3D CNN adds to memory usage which is counter-intuitive to saving memory, judging from the results in Table 3, a small 3D CNN like Simple3D with only 3K parameters can already provide decent features. If an unmodified video transformer that solely relies on attention is used, considerable data pre-processing should be done to address spatial redundancy for facial emotion recognition.

3D CNNs: Both 3D CNN models show promising results in terms of accuracy. I3D gave the most accurate predictions, but the unimodal vision I3D model suprisingly outperformed the multimodal I3D variant. This is believed to be due to ImageNet pretraining being more suited for unimodal vision. Simple3D also had decent results in unimodal audio and multimodal despite being a tiny model with only 3262 parameters and no pretraining.

Furthermore, both models were able to learn multimodal interactions. This is especially true for the small Simple3D model which greatly benefited from this in the multimodal experiment. This could be attributed to dividing mel spectrograms along the time dimension to temporally align vision and audio. Converting audio to 3D might be a waste of weights in some scenarios, but it does help with multimodal interaction.

General Discussion: Although 3D models show promising results, they have a caveat. They require 3D data which comes with expensive computational resources. This is expanded upon in the next section. As a result, all 3D models had little to no hyperparameter tuning in these experiments which would have only improved results.

Another thing of note is the ablated results. Agnosticity was not the focus, but we briefly mention it here. I3D showed that it can reasonably handle missing modalities. It offers similar accuracies to the unimodal ViT trained from scratch when vision is masked (Table 1) and when audio is masked (Table 2). Simple3D exhibits similar agnostic properties, but it does not produce comparable results to other models like I3D does.

5.8 2D vs 3D

I3D offers either the best or competitive performance across all unimodal and multimodal experiments (Tables 1, 2, 3), but it may not be suitable depending on the task. 3D data inherently comes with more costs compared to 2D such as longer training / inference times and higher memory. When resources are not an issue, 3D encoders can offer better performance. But in situations where resources are limited, 2D encoders may be the better option. We look at vision data for comparison. Video data is downsampled to a single frame in 2D data compared to extracting frames per second in 3D data. In addition to this, 3D data requires padding if there are unequal number of frames per video. Both reasons enable 2D encoders to have significantly faster processing times and lower memory constraints.

PT ViT had slightly better unimodal audio and multimodal accuracies compared to I3D, but fell behind in unimodal vision. However, it is a 2D architecture and does not have the computational costs that come with 3D data. Therefore, PT ViT may be the better option in general. Furthermore, 2D vision transformers are more suitable for emotion recognition compared to 3D vision transformers. 2D vision transformers performed better and had lower computational complexities, whereas the 3D vision transformers suffered from spatial redundancy and had higher computational costs.

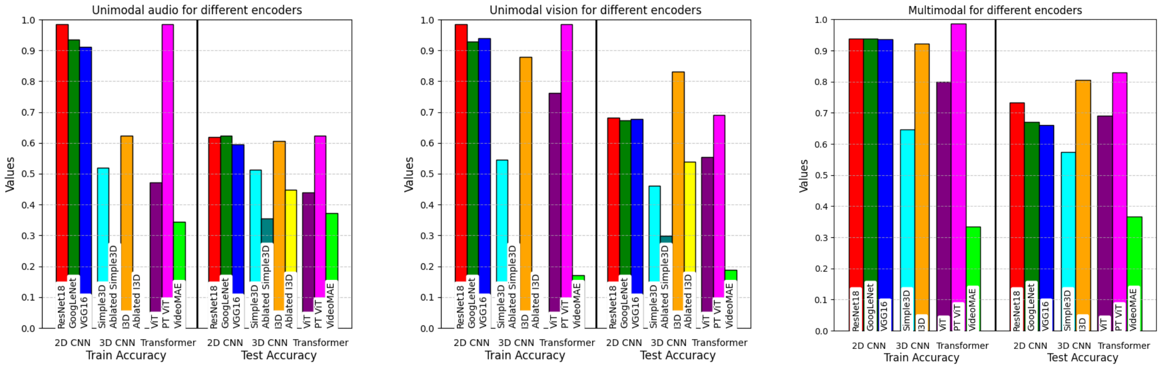
6 Conclusion & Future Work

We have successfully implemented the unimodal and multimodal audio and vision pipelines with 2D CNN, 3D CNN, ViT, and VideoMAE as encoders. We tested our pipelines on a fullscale version of CREMA-D containing 7442 samples, and all the 6 emotion classes. For 2D video pipelines, we performed hyperparameter tuning (in a non-exhaustive manner), and identified the best modes of operation. We compared 2D and 3D pipelines against each other in terms of their test accuracies, reasoned the observed behavior, and analyzed the implications.

For 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. This, we believe, will improve the performance of our pipelines with ViT significantly. In the 3D pipeline, adding a small 3D CNN may help mitigate spatial redundancy in videos and also address joint-space attention memory constraints in VideoMAE. Finally, 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.

7 Contributions

- Anuroop
  1. Implemented unimodal audio, vision, and multimodal pipelines with ResNet18. Other 2D CNN pipelines were modelled on this pipeline.
  2. Implemented unimodal audio, vision, and multimodal pipelines with ViT. Performed full-scale experiments for unimodal audio and multimodal modalities on this pipeline.
  3. Implemented 2D data pre-processing scripts for full-scale experiments that included spectrogram generation and storage of pre-processed data in the form of `.npy` files to avoid running pre-processing stage for each experiment.
  4. Implemented barplot generation script.
  5. Fixed bugs in 2D data pre-processing that increased performance in audio modality especially.
  6. Major contribution in preparing midterm and final presentation decks.
  7. Final report - Sections 2, 5.1 - 5.4, part of 6.
- Riya
  1. Implemented unimodal audio, vision, and multimodal pipelines for GoogLeNet and RGB images. Conducted full-scale experiments for this pipeline with hyperparameter tuning on learning rates and batch sizes.
  2. Performed hyperparameter tuning on full-scale ViT pipeline involving batch size, heads, blocks, and dropout rates.
  3. Customized ViT pipeline for RGB images on pre-trained ViT (PT ViT) from HuggingFace. Conducted full-scale experiments for unimodal audio, vision, and multimodal for this pipeline.
  4. Identified bugs in 2D experiments that helped fix faulty 2D performances and



(a) Unimodal audio

(b) Unimodal vision

(c) Multimodal

Figure 2: Comparison of accuracies of encoders for different modalities

Type	**Encoder**	**Train Loss**	**Train Acc.**	**Test Loss**	**Test Acc.**
2D CNN	ResNet18	1.0602	0.9848	1.4199	0.6182
	GoogLeNet	1.1124	0.9344	1.4152	0.6244
	VGG16	1.1326	0.9113	1.4438	0.5962
3D CNN	Simple3D	1.213	0.519	1.294	0.514
	Ablated Simple3D	-	-	1.792	0.354
	I3D	0.991	0.623	1.022	0.605
	Ablated I3D	-	-	2.707	0.448
Transformer	ViT	1.3354	0.4710	1.4487	0.4401
	Pretrained ViT	0.0545	0.9846	1.7709	0.6231
	VideoMAE	1.566	0.344	1.512	0.372

Table 1: Unimodal audio metrics for different encoders

Type	**Encoder**	**Train Loss**	**Train Acc.**	**Test Loss**	**Test Acc.**
2D CNN	ResNet18	1.0603	0.9848	1.3597	0.6815
	GoogLeNet	1.1169	0.9275	1.3649	0.6742
	VGG16	1.1049	0.9404	1.3641	0.6765
3D CNN	Simple3D	1.160	0.546	1.417	0.462
	Ablated Simple3D	-	-	2.074	0.298
	I3D	0.334	0.878	0.463	0.831
	Ablated I3D	-	-	1.928	0.540
Transformer	ViT	0.6374	0.7612	1.6227	0.5536
	Pretrained ViT	0.0453	0.9857	1.3825	0.6895
	VideoMAE	1.795	0.170	1.790	0.188

Table 2: Unimodal vision metrics for different encoders

Type	**Encoder**	**Train Loss**	**Train Acc.**	**Test Loss**	**Test Acc.**
2D CNN	ResNet18	1.1050	0.9379	1.3074	0.7326
	GoogLeNet	1.1069	0.9390	1.3678	0.6711
	VGG16	1.1081	0.9356	1.3785	0.6608
3D CNN	Simple3D	0.918	0.647	1.169	0.573
	I3D	0.211	0.923	0.629	0.806
Transformer	ViT	0.5566	0.8002	1.0465	0.6909
	Pretrained ViT	0.0397	0.9867	0.7543	0.8290
	VideoMAE	1.575	0.334	1.513	0.366

Table 3: Multimodal metrics for different encoders

Class	Encoder	# (Train + test)	bs	lr	optim	Loss	ep	dr	GPU
3D CNN	Simple3D(V)	$5951 + 1488$	8	0.001	Adam	CE	27	0	A100 40GB
	Simple3D(A)						36		
	Simple3D(M)								
	I3D(A)						5	0.5	
	I3D(V)						9		
	I3D(M)		4				16		
2D CNN	ResNet18(A)	$5200 + 2229$	32	0.0001			50	-	P100
	ResNet18(V)								
	ResNet18(M)			0.001					
	GoogLeNet(A)		64	0.0001					
	GoogLeNet(V)								
	GoogLeNet(M)								
	VGG16(A)		16	0.00001					
	VGG16(V)								
	VGG16(M)		32						
Transformer	ViT(A)		16	0.0001				0.4	
	ViT(V)								
	ViT(M)								
	Pretrained ViT(A)		32				16	-	
	Pretrained ViT(V)								
	Pretrained ViT(M)						21		
	VideoMAE(A)	$5951 + 1488$	8	0.001			3	0.5	A100 80GB
	VideoMAE(V)						4		
	VideoMAE(M)								

Table 4: Training setup

executed a rerun of the 2D experiments for hyperparameter-tuned GoogLeNet, ViT, and PT ViT with corrected notebook.

5. Midterm and final report - Sections 3, part of 6.

- Aashi
    1. Implemented and fine-tuned unimodal audio, vision, and multimodal pipelines on subset data with VGG16.
    2. Converted grayscale data to RGB images for implementation. Further fine-tuned the VGG16 model with best-fit results, analysis, and discussions.
    3. Implemented and fine-tuned unimodal audio, vision, and multimodal on full-scale data pipelines with VGG16. Converted grayscale to RGB and to fit the NumPy arrays. Modelled those with fine-tuning based on parameter changes in learning rates and batch sizes.
    4. Executed and fine-tuned the unimodal audio, vision, and multimodal pipelines on full-scale data with ResNet18 pre-trained model to get the best possible results.
    5. Executed a rerun of the 2D experiments for ViT, ResNet18, and VGG16 with corrected notebook.
    6. Designed architecture diagram (Figure 1) for the project in final report on the basis of Figure 3 in [Liang et al., 2023].
    7. Midterm and final report - Sections 1, 4.
- Wilson
    1. Solely responsible for all of 3D parts including: 3D data pre-processing, 3D encoder training/testing (unimodal and multimodal), 3D analysis and discussion, and modality ablations on CNN models.
    2. Performed brief data cleaning on the raw CREMA-D dataset which identified the 3 recording errors.
    3. Identified bugs in 2D experiments that helped fix faulty 2D performances.
    4. Final report - Sections 2, 5.5 - 5.8.

All team members have actively contributed to the project. Furthermore, everyone contributed to proof-reading both the presentation decks and the report.

8 Miscellaenous

Our experiments are available as `.ipynb` notebooks and `.py` scripts accompanied with a README file and can be reproduced. Code-base is hosted on this GitHub repository – https://github.com/ksanu1998/multimodal_course_project. Other experiment resources are stored here – https://drive.google.com/drive/folders/1BhpgUDgbYwoTaTO6Yo8M3uR0Clw0bkiC?usp=sharing and https://drive.google.com/drive/folders/1Q1LFiq2KZPyYTuEJhbQY38uu9FE0Jl-g?usp=sharing. I3D PyTorch implementation was taken from https://github.com/piergiaj/pytorch-i3d/tree/master with a small modification to the forward method. ViT PyTorch implementation was adapted from https://theaisummer.com/vision-transformer/.

References

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Sai Srujana Buddi, Utkarsh Oggy Sarawgi, Tashweena Heeramun, Karan Sawnhey, Ed Yanosik, Saravana Rathinam, and Saurabh Adya. Efficient multimodal neural networks for trigger-less voice assistants, 2023.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Bin Jiang, Qiang Ren, Fei Dai, Jian Xiong, Jie Yang, and Guan Gui. Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method. In Qilian Liang, Xin Liu, Zhenyu Na, Wei Wang, Jiasong Mu, and Baoju Zhang, editors, *Communications, Signal Processing, and Systems*, pages 59–66, Singapore, 2020. Springer Singapore.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

Yuanyuan Lei and Houwei Cao. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969, 2023. doi: 10.1109/TAFFC.2023.3234777.

Yu-Jhe Li, Jinhyung Park, Matthew O'Toole, and Kris Kitani. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2022.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning, 2023.

Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapes. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20, 2023. ISSN 1939-3539. doi: 10.1109/tpami.2023.3243465. URL http://dx.doi.org/10.1109/TPAMI.2023.3243465.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022b.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Yufeng Yin, Jiashu Xu, Tianxin Zu, and Mohammad Soleymani. X-norm: Exchanging normalization parameters for bimodal fusion. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 605–614, 2022.