First Model

2D Parallel CNN

Original model Using RAVDESS Dataset

Model	Augmentation	Learning rate	Extracted features	Dataset	Accuracy	Number of emotions
2D parallel CNN with four transformer layer (self attention layer)	White gaussian noise on train & vaild and test	0.01	MFCC	Speech	64.87%	8
2D parallel CNN with four transformer layer (self attention layer)	White gaussian noise on train & vaild and test	0.01	MFCC	Speech & song	73%	8

Our Models

Model	Augmentation	Learning rate	Extracted features	Dataset	Accuracy	Number of emotions
Same original	White gaussian noise on train & vaild and test	0.01	MFCC	Speech & Song	76%	6
change in dropout ratio	white gaussian noise on train only	0.01	MFCC	Speech & Song	79.9%	6
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	79.4%	6
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	84.96%	4
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	80.12%	5
Same original	White gaussian noise on train & vaild and test	0.01	MFCC	Speech & Song	72%	6
3 parallel CNN Blocks	white gaussian noise on train only	0.01	MFCC	Speech & Song	77.51%	6
2 transformer layers instead of 4	white gaussian noise on train only	0.01	MFCC	Speech & Song	77.99%	6
1 Block CNN Instead of 2D Parallel CNN	white gaussian noise on train only	0.01	MFCC	Speech & Song	77.03%	6
Change only in Drop ratio	white gaussian noise on train only	0.01	MFCC	Speech & Song	77%	6

Adding layer to original model (4 layers in each Block)	white gaussian noise on train only	0.01	MFCC	Speech & Song	75.12%	6
Same original	white gaussian noise on train only	0.001 & weight decay = 1e-6	MFCC	Speech & Song	71.62%	7
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	73.36%	7
Same original	Without augmentation	0.01	MFCC	Speech & Song	72.05%	7
Same original	White gaussian noise on train & vaild and test	0.001 & weight decay = 1e-6	MFCC	Speech & Song & Dataset 60,20,20	72.57%	7
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	75%	6
Same original	White gaussian noise on train & vaild and test	0.01	MFCC	Speech	68%	7
Same original	white gaussian noise on train only	0.01	MFCC	Speech & Song	78.95%	5
Drop second CNN layer by 0.8 ratio	white gaussian noise on train only	0.01	MFCC	Speech & Song	69.86%	6
Same original	white gaussian noise on train only	Lr Decrease with time	MFCC	Speech & Song	80.86%	6

Second Model

CNN

Dataset	Original			dropout			New layer			Original + Augmentation		
	worst	best	avg	worst	best	avg	worst	best	avg	worst	best	avg
RAVDESS	69.95%	77.66%	72.93% (+/- 2.61%)	72.95%	79.23%	76.02%	70.7%	78.99%	76.12% (+/- 2.85%)	79.23%	83.57%	81.45%
SAVEE	61.11 %	75.00%	70.56% (+/5.08%)	66.67%	80.56 %	72.78% (+/- 5.02%)	65.28%	79.17%	75.56% (+/- 5.24%)	61.11%	77.78%	71.11%
RAVDESS + SAVEE	64.02%	70.93%	68.00% (+/- 2.76%)	62.68%	66.05%	62.85% (+/- 2.66%)	70.99%	73.40%	72.65% (+/- 0.85%)			
Kfolds=5												

Third Model

CNN

Model	Dataset	Feature extraction	Epoch	Batch size	Emotions	Accuracy
CNN	RAVDESS Speech	MFCC	500	20	4	61
						60
						65
	RAVDESS Speech & Song		500	20	4	75
			500	10	4	77
			500	5	4	73
			500	4	6	75
			200	24	6	74
			200	26	6	75
		MFCC & Mel &chroma	500	10	6	81

Conclusion

The highest accuracy we achieved is on the second model using Batch normalization and also using augmentation.

Dataset	MODEL	AVG Accuracy	Best Accuracy	Worst Accuracy
RAVDESS Speech & song	CNN	83.13%	86.47%	81%
SAVEE	CNN		72.22%	69%