Master thesis

Emotion Recognition in videos using Multi-modal Neural Network

Himanshu Gupta
Automatic Control and Robotics and specialization Robotics
Academic year 2018/2019

Supervisor: prof. Dr. inż. Andrzej Kordecki

1. Introduction

In recent years, there has been a rise in social robotics which require robots to communicate with humans, and for communicating with humans these robots need to recognise emotions to understand the human intent and provide the correct response. For recognizing the emotions, we used multi-modal neural networks because humans' convey emotions through facial expressions, voice and body gestures. In this work, voice and facial expressions are used for classification. We had considered only 7 basic emotions (neutral, happy, sad, fear, angry, surprise and disgust). The aim was to evaluate various deep learning architecture to see which network is best suited for a robotic application.

2. Methodology

In this section, we discuss about the dataset used, preprocessing of the data, different network architecture used for classification, and training method.

- **Dataset Description:** Three datasets are used, RAVDESS Dataset: for training the networks, RML dataset: to check the robustness of the networks, and Affectnet dataset: for training the CNN used in LRCN networks.
- **Preprocessing:** In case of audio preprocessing, the audio is extracted as wav file, OpenSMILE is used for features extraction and normalized using SKlearn's scale function. In case of video preprocessing, video is read frame-by-frame, faces are detected and cropped and saved as JPEG files using OpenCV. For temporal information, optical flow is calculated using OpenCV's DTVL-2 function .
- **Implementation Framework:** Python language is used for implementation. Network modelling is done using Keras library with tensorflow backend. The networks are trained in Google's colab environment which has Nvidia's Tesla GPU of 16 GB.
- **Network Architecture:** For classification using audio, SVM and 1D CNN classifiers are used. For classification using video, ResNet50, C3D and LRCN model are used. For classification using optical flow, ResNet50 and CNN similar to AlexNet is used. The multi-modal network is made by combining the results from audio and video CNN by i) averaging and ii) SVM classifier.
- **Training Parameters:** For training the networks, cross-entropy loss is used and optimized using RMSProp with learning parameter 0.001. Other Keras functions are used to support the training, TensorBoard, ModelCheckpoint and EarlyStopping.

3. Results

Table: Comparison of training time and validation accuracy for SVM classifier and CNN model on audio dataset.

	Emobase 2009			Emobase 2010			Emobase Large		
	Train-acc	val-acc	training time (sec)	Train-acc	val-acc	training time (sec)	Train-acc	val-acc	training time (sec)
SVM	-	0.49	22	-	0.67	57	-	0.48	426
CNN	0.963	0.487	689	0.962	0.656	1116	0.88	0.48	4720

Table: Comparison of training for three Video CNN models.

	ResNet50				C3D				LRCN (ResNext50)			
	train acc	train loss	val acc	val loss	train acc	train loss	val acc	val loss	train acc	train loss	val acc	val loss
Metric	0.95	0.16	0.74	1.3	0.89	0.3	0.76	0.8	0.4	1.6	0.45	1.6
Time(s)/epoch	33				176				200			
Parameters (M)	2.5				0.33				2.1			

Figure: CNN Optical flow results

Table: Comparison of accuracy and loss for Multi-modal CNN.

	RAVDESS Dataset				RML Dataset
	train acc	train loss	val acc	val loss	acc
ResNet+Audio CNN+Avg	-	-	0.77	-	0.37
C3D+Audio CNN+Avg	-	-	0.74	-	0.38
LRCN+Audio CNN+Avg	-	-	0.66	-	0.4
ResNet+Audio CNN+SVM	0.983	0.38	0.725	0.4	0.4
C3D+Audio CNN+SVM	0.984	0.38	0.73	0.4	0.4
LRCN+Audio CNN+SVM	0.98	0.38	0.656	0.4	0.4

4. Conclusion

- Emobase 2010 audio feature set gave best accuracy (0.66)
- For video classification, ResNet50 and C3D gave same accuracy (0.76) on RAVDESS dataset.
- For robotic application, the LRCN model is well suited, robust to image conditions (background, gender, age, and noise).
- C3D can also be used in robotics application given its compact size (0.33M parameters)
- CNN trained on optical flow data did not train maybe due to sparsity in the data.

5. Future Work

- For improving the accuracy of LRCN, a dataset that has more equally distributed images for each emotion category should be used. Other CNN architecture like Xception, VGG16, VGG19, InceptionV3 should be considered like to get the right balance of network size and accuracy.
- Evaluate combination of the activation functions, batchnorm layer placement with different optimizers to see for the improvements in the network's accuracy.