Chapter 6

Sattriya Dance Gesture Classification from Video with Convolutional Neural Network

Deep learning thrives in various areas, from gesture recognition to video classification, human activity recognition, speech recognition, and labeling. Since 2010, learned-based features have become powerful than handcrafted features. The arrival of deep learning technology changes the scenario of machine learning. Many works using traditional machine learning methods move around to deep learning approaches to resolve multi-class classification problems.

In this chapter, we propose a deep learning approach for ground exercise classification. Convolutional Neural Network is used for learning the features automatically. Transfer learning is applied in this model. Here we have selected VGG16 architecture which is pre-trained with Imagenet. Two dense layers are included additionally with this model. Softmax is used as activation function for this multi-class classification problem.

The rest of the chapter is organized as follows: A brief description of the need of deep learning approach is discussed in Section 6.1. Section 6.2 describes the technical details of deep learning approach. Section 6.3 presents the proposed method. In Section 6.4, experimental results are discussed. Finally, in Section 6.5, the summary of this chapter with the scope of future work is concerned.

6.1 Why Deep Learning

Deep learning techniques have shined the world of machine learning. It helps in learning complex features automatically. Today, learned-based features become powerful than handcrafted features in classification task. Generally, in computer vision, inputs are often images. The convolutional layer is used which processes grid shape inputs. In deep learning approaches, datasets of small size led to over fitting during the training phase. Transfer learning can be used to resolve the over fitting problem.

6.2 Technical Details of Deep Learning Elements

The essential elements of deep learning for dynamic gesture recognition are presented:

- For multi-class classification problem, softmax can be used as activation function. It gives a class-conditional probability vector as output.
- We use Convolutional Neural Networks to handle dynamic aspects of videos.
- Deep learning algorithms with small size dataset led to over fitting problem at the time of training phase. Transfer learning can be used to resolve the problem of over fitting that uses another similar dataset.

6.2.1 Softmax function

For multi-class classification problem, softmax activation function can be used. The output of a classification task is a categorical variable. A categorical variable has a determined number of possible and discrete events. For a classification problem with two possible events, e.g., true or false can be referred to as binary variables. Again, in a seed classification problem from images, the possible classes could be different kinds of seeds in the dataset. Each sample in the dataset is assigned to one of those finite categories. They differ from quantitative variables in that the distances from one category to another are equals, regardless of the number of classes. A single scalar can be used to represent the outputs, but the distances between each class would not be equal. To fix this issue, deep learning models designed for classification use one-hot encoding to represent their outputs. It consists of a vector in which the size is equal to the number of categories, fill with 0 and a 1 in the cell of the category to which the input belongs. This encoding scheme can be used as a particular stochastic vector that represents the probability that an input belongs to each class, called a class-conditional probability vector.

A model needs two elements to output a stochastic vector. First, to obtain the size of the number of classes we need the last layer of the model. Second, softmax activation function is used in the previous layer to compute a class-conditional probability vector. The output of the softmax function is a categorical probability distribution that indicates the probability that the input belongs to any of the categories.

6.2.2 Cross-entropy cost function

A key strand of deep neural network training is the choice of the cost function. There are various cost functions that can be used in a classification task. Cross-entropy is a commonly used cost function for classification. Cross-entropy is a measure to calculate the distance between two probability distributions. In a classification problem with multiple classes, cross-entropy cost function is normally used.

6.2.3 Convolutional Neural Networks (CNN)

Convolutional Neural Networks are used to learn features automatically from the dataset. CNNs are a kind of neural network for processing grid-like data. It includes time-series vectors, which are grid-like data when concatenated, and, of course, images which are 2D grids of pixels. 2D image frames are obtained from the videos which are the input for this gesture classification problem. As the image frames are used as input to this system, CNN architecture is introduced to resolve the problem.

In traditional neural network layers such as multilayer perceptron, every output interacts with every input. The number of parameters of a neural network model is proportional to its input size. When processing an image, the input might have thousands or millions of values and a conventional multilayer perceptron will see its number of parameter and runtime explode. Also, such network architecture does not take into account the spatial structure of the image. By handling images as vector of pixels, it does not allow the network to benefit of the strong spatially local correlation present in images which are important features in recognition task.

The goal in a classification process is to propose a classifier which works correctly on new previously unseen inputs. The ability to handle unobserved inputs is called generalization. Typically, a dataset is composed of two non-overlapping sets- the training set and the test set. The training set is a subset of the dataset based on which the model learns the features. On the contrary, the test set is comprising the unknown data during the training phase of the model. The classifier has to reduce error measures between its outputs and ground-truths through an optimization procedure. This error measure computed on the training set during training phase is called training error and generalization or test error when computed on the test set. The effectiveness of a learning algorithm is determined by its ability to make the training error small and to reduce the difference between the training and the test error called generalization gap.

6.2.4 Transfer Learning

Transfer learning is a machine learning technique where relevant part of a pre-trained model is taken and apply it in a similar problem. The lack of data on specific tasks is one of the main reasons to use it, since collecting and labelling data can be very expensive and can take time, and recent concerns with privacy make difficult to use real data from users. The use of transfer learning helps to fast prototype new machine learning models using pre-trained models from a source task since training on millions of images can take time and requires expensive GPUs.

6.3 Overview of the Proposed Framework

The proposed framework for deep learning approach for ground exercise classification is shown in Figure 6.1.

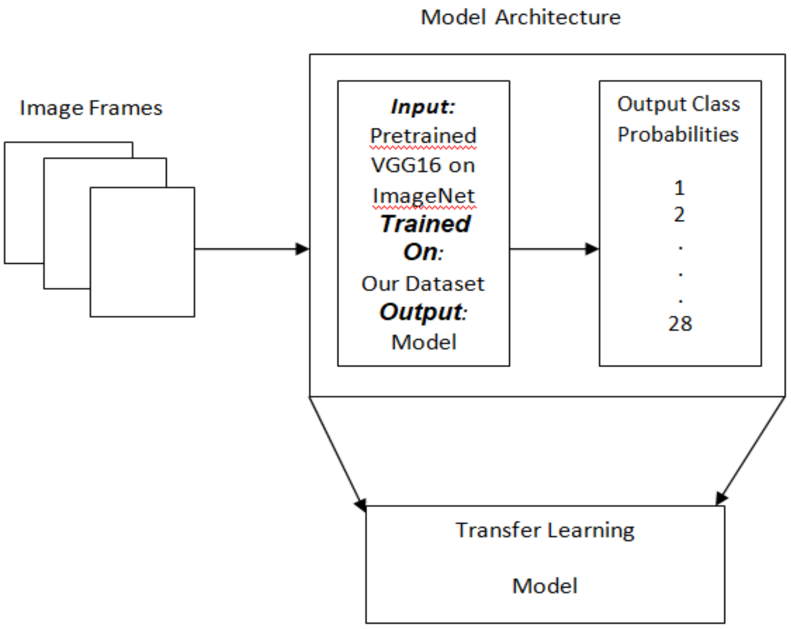

Figure 6.1 Overview of the proposed framework

This work uses the transfer learning concept in CNNs to achieve better performance in classification. The hyper parameters such as number of epochs, optimizer and batch size were optimized using grid search. Keras ships out-of-the-box with five Convolutional Neural Networks that have been pre-trained on the ImageNet dataset: VGG16, VGG19, ResNet50, Inception V3 and Xception. From the experiments we have selected the model as an imagenet pre-trained VGG16 architecture.

6.4 Experimental Results

Our proposed system has been implemented with keras, the deep learning toolbox for python, using tensorflow as the backend.

6.4.1 Dataset Description

The data set used in our work comprises of 560 videos of 28 ground exercises, i.e. classes. A total of 368 videos are considered for training purposes. The remaining videos in each class are considered for testing.

6.4.2 Results and Discussion

From the experiments we have selected the VGG16 architecture that is pre-trained with imagenet. It has 2 additional dense layers and softmax is used as activation function for our 28 classes. Model has an overall classification rate of 95.52%. Imagenet is used as pre-training dataset and our own created dataset is used for training. A better classification rate is achieved using this architecture.

6.5 Summary

Learning features on images is a complex task. Besides, CNN architecture contains lot of parameters and, so, is not usually trained from scratch with random initialization. Mainly transfer learning depends on the size of the source dataset and its similarity to the original dataset. Transfer learning is used in this CNN architecture to achieve better classification. Grid search is used to tune the related hyper parameters. Our model has achieved overall classification rate of 95.52%.