ED-NET: Educational Teaching Video Classification Network*

Anmol Gautam[1], Sohini Hazra[2], Rishabh Verma[2], Pallab Maji[2], and Bunil Kumar Balabantaray[1]

[1] National Institute of Technology Meghalaya, Shillong, India
{t20cs004,bunil}@nitm.ac.in
[2] GahanAI, Bengaluru, India
https://gahanai.com/
{sohini.hazra,rishabh.verma,pallab.maji}@gahanai.com

Abstract. Convolutional Neural Networks (CNNs) are widely used in Computer Vision-based problems. Video and image data are dominating the internet. This has led to extensive use of Deep Learning (DL) based models in solving tasks like image recognition, image segmentation, video classification, etc. Encouraged by the enhanced performance of CNNs, we have developed ED-NET in order to classify videos as teaching videos or non-teaching videos. Along with the model, we have developed a novel data set, Teach-VID, containing teaching videos. The data is collected through our e-learning platform Gyaan, an online end-to-end teaching platform developed by our organization, GahanAI. The purpose is to make sure we can restrict non-teaching videos from being played on our portal. The models proposed along with the dataset provide benchmarking results. There are two models presented one that makes use of 3D-CNN and the other uses 2D-CNN and LSTM. The results suggest that the models can be used in real-time settings. The model based on 3D-CNN has reached an accuracy of 98.87% and the model based on 2D-CNN has reached an accuracy of 96.34%. The loss graph of both models suggests that there is no issue of overfitting and underfitting. The proposed model and data set can provide useful results in the field of video classification regarding teaching vs non-teaching videos.

Keywords: Video Classification · Deep Learning · CNN · LSTM.

1 Introduction

The amount of audio-video data has increased exponentially in recent times. This has led researchers to develop methods and data sets working directly in video modality. Video classification has seen strong progress in the recent past. Advancements in Deep Learning (DL) based models and techniques have led to the introduction of powerful models that have solved video-based vision problems. But model architecture development is only part of the problem. To develop

* Supported by organization GahanAI, Bengaluru, India

a solution, a data set for solving the specific problem is necessary. Large-sized data sets like Image-net were key in the development of DL-based solutions for image-related tasks.

Currently, Convolutional Neural Networks (CNNs) based models have been extensively used for video classification. The information in video data has a Spatio-temporal aspect along with sequential information. This also makes handling video data directly a challenging task. The video is composed of sequential images with overlapping contents.

Due to a shift in teaching mode, in the wake of the global pandemic. We have witnessed a new form of data modality that started dominating the internet. The online shift of teaching institutions ranging from schools to graduate colleges. Though, YouTube and other related platforms had online content related to education and training categories. But this recent shift has led to entire classes being held on platforms like Google Meet, Microsoft Teams, and some custom-made applications like the e-learning portal developed by our organization. A new problem that has emerged in online teaching mode is whether the video being played on the platform belongs to the teaching category or not. In this regard, we have developed a system that can classify the video being played on our e-learning platform as teaching or not. If the video is categorized as a Non-Teaching category then the stream will stop immediately.

In this paper, we introduce Teach-VID, a data set containing online teaching videos which will be useful for video classification tasks. In our study, we have treated video classification as producing relevant labels given video clips. Some frames of a video clip from the Teach-VID data-set can be seen in Figure 1.

Along with the data set, we have also proposed a model that can classify the given video as teaching or non-teaching. The baseline model can be used for benchmarking purposes. Our contributions can be summarized as follows:

- Teach-VID data-set, a novel data-set containing online video clips. Each clip is of sixty-second duration and is extracted from the videos uploaded on our e-learning platform.
- We propose two models for the video classification task. One is using 3D convolutional layers for feature extraction and Dense layers for classification. Another model is developed using 2D convolutional layers for frame-wise feature extraction and using the LSTM layer to incorporate sequential information. In our study, both models have shown competitive results and can be used as bench-marking models for further work on the Teach-VID dataset.

2 Related Work

The task of video analysis is computationally expensive due to complicated data modalities. The sequence information has to be processed efficiently to get the most out of it. In this regard, Bhardwaj et al. [2] have proposed a Teacher-Student architecture that is computationally efficient. They utilized the idea of knowledge distillation. The network makes use of a fraction of frames at inferencing time. the student network performs on a similar level to the teacher network on the Youtube-8M [1] dataset. Xu et al. [15] have utilized networks with two streams that learn static and motion information. They have utilized the late fusion approach in order to reduce overfitting risk. Peng et al. [10] have used spatiotemporal attention along with spatial attention to obtain discriminative features to achieve better results. Appearance and Relation Network was introduced by Wang et al. [14] to learn video representation in an end-to-end manner. The SMART block introduced in the network separates out Spatio-temporal information in spatial modeling and temporal modeling. Tran et al [13] have developed Channel Separated Convolutional Network which factorizes 3D convolutions. this helps to reduce the computation along with boosting the results. Long et al [8] have used the idea of using pure attention networks for video classification. They have proposed Pyramid-Pyramid Attention Clusters that incorporates channel attention and temporal attention. Pouyanfar et al [11] have extracted Spatio-temporal features from video sequences using residual attention-based bidirectional Long Short-Term Memory. Further, to handle data imbalance they have used weighted SVM. NeXtVLAD is introduced by Lin et al [7], to aggregate frame-wise features into a compact feature vector. They decompose high dimensional features into a group of smaller dimensional vectors. The model was found to be efficient in aggregating temporal information. Li et al [6] have proposed SmallBig Net that uses two branches to incorporate core semantics through one and contextual semantics through another.

3 Dataset Development and Description

The Teach-VID dataset contains 60-second clips of teaching videos uploaded on our e-learning platform developed by our organization. Our web-based product is an e-learning platform that aims to provide an online medium to conduct educational and e-learning activities. this platform is available to institutes, coaching, schools, and individuals who want to create and manage classes in an online manner. The video uploaded on our platform is used to create the dataset. Each video was pre-processed to remove identification markers like visual clues and audio clues to ensure privacy.

To train the models for the classification task we have used the entire dataset of 746 clips of 60 seconds along with teaching videos taken from platforms like YouTube. We are also releasing a smaller version of the dataset to be used on our website after making sure the faces of either students or teachers are removed due to privacy concerns. The total number of clips that is released are 58 and the audio stream is removed from them. The dataset can be accessed at the link `https://gahanai.com/dataset.php`. The frames of the clips have been resized to the dimension of (256,256,3). In order to make the model more robust further collection of data is ongoing. To make sure we had enough diversity we trained the model by taking videos of teaching categories from open platforms like YouTube `www.youtube.com` .The remaining videos will be released in due course of time as they contained the faces of teachers or students. In order to ensure the anonymity and privacy of the users of our portal, the said video has been kept private. After taking the necessary steps to ensure privacy the remaining videos will be released to be used openly.

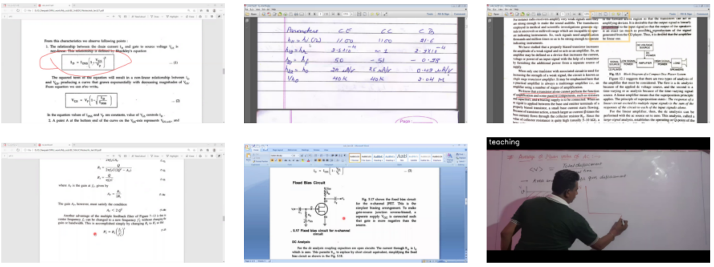

Fig. 1. Sample Frames from Teach-VID Dataset

4 Methodology

This section focuses on the proposed ED-Net model and its components used for the video classification task. The methodology we followed consists of several steps which are discussed herewith. First, we extracted twenty frames from each clip of the Teach-VID dataset and from UCF-50 dataset [12] as well. This allowed us to represent a given clip in a four-dimensional NumPy array. This is done for all the clips. As it is a binary classification problem, we formed two classes ’teaching’ and ’non-teaching’. Now our two proposed models perform classification using these labeled training data. In the first method which makes use of 3D-CNN, the architecture itself takes into consideration the spatial and temporal information, as discussed in the following section. In the second model, we have used 2D-CNN to first extract features from each frame, therefore for 20 frames per video clip, we extract 20 feature maps and two LSTM layers that combine the temporal information present. Therefore, in the first method spatial and temporal information is processed simultaneously and in the second, first, we extract spatial features and then temporal features. Finally, a classifier is used to classify the extracted features.

4.1 Problem Formulation

Let $X \in \mathbb{R}^{h \times i \times j \times k}$ is input video and Y is its corresponding label. Y belongs to binary class i.e. teaching videos will be encoded as label ’0’ and non-teaching videos will be labeled as ’1’. The proposed model after training classifies each given video clip as belonging to Class 0 or Class 1.

4.2 Model Architectures

We have proposed two different architectures to explore the usability and efficacy of the task of teaching video classification. The proposed model, ED-Net Model-A, uses 3D convolution to take volumetric data and use it to extract temporal and spatial features in an end-to-end manner. Another model, ED-Net Model-B, uses 2D convolutional layers first to extract features from each frame and finally uses an LSTM layer to incorporate sequential information. Both models have performed well on the task at hand. A detailed description of each model is given in the following sections. The goal behind developing simple models is to give bench-marking results on the data set at hand for the task of video classification.

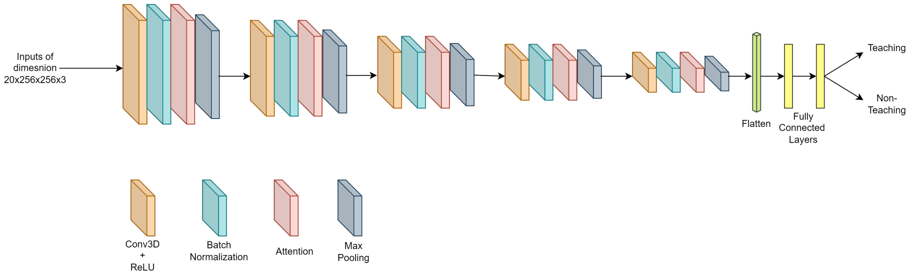

Fig. 2. Proposed Model Architecture using 3D-CNN

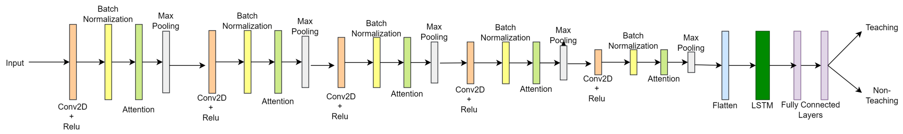

Fig. 3. Proposed Model Architecture using 2D-CNN and LSTM

Attention Module We have used Triplet Attention (TA) [9] and a modified version of it which can be used along with 3D-CNN, the module provides to refine the feature map representation by adding little cost to the network. Triplet attention improves the overall performance of the model as the ablation study done by authors [9] has established this. The TA module uses three paths to calculate the interaction between the three dimensions, height, width, and channel. The three paths are represented as (HxWxC), (HxCXW), and (WxHxC) where the Input tensor is represented as (HxWxC).

Each path is represented as O_1, O_2 and O_3.

X is the input tensor and X' is the output tensor.

$f_{perm}(.)$ represents permutation of the tensor along an axis. $f_{BN}(.)$ represents Batch Normalization (BN), $f_{k\times k}(.)$ is $k \times k$ convolution and $f_{concat}(.)$ is concatenation operation on the input tensors along respective axis. σ represents sigmoid activation function. We have placed the BN layer to avoid internal covariate shift due to three paths in the triplets.

$$O_1 = \sigma\left(f_{BN}\left(f^{7\times 7}\left(f_{concat}\left(\sqsubset F^S_{avg}, F^S_{max} \sqsupset\right)\right)\right)\right) \tag{1}$$

$$O_1{}' = X \odot O_1 \tag{2}$$

$$O_2 = \sigma\left(f_{BN}\left(f^{7\times 7}\left(f_{concat}\left(\sqsubset F'_{avg}, F'_{max} \sqsupset\right)\right)\right)\right) \tag{3}$$

$$O_2{}' = X' \odot O_2 \tag{4}$$

$$O_3 = \sigma\left(f_{BN}\left(f^{7\times 7}\left(f_{concat}\left(\sqsubset F''_{avg}, F''_{max} \sqsupset\right)\right)\right)\right) \tag{5}$$

$$O_3{}' = X' \odot O_3 \tag{6}$$

$$X' = X \odot \left(f_{BN}\left(f_{1\times 1}\left(\sqsubset O_1'; O_2'; O_3' \sqsupset\right)\right)\right) \tag{7}$$

ED-Net Model-A (3D-CNN based model) The proposed model takes M frames of NxNx3 dimensions as input and outputs a corresponding label value. The 3D-CNN-based architecture using 3D convolutions is gaining popularity to extract features from video. They use 3D kernels that slide along three dimensions to extract spatial and temporal features. The architecture of the proposed model is composed of the following layers:

- 3D-Convolution Layer: This layer makes use of 3D convolutions or filters which slide across the 3-axis to extract low-level features. The output of the filter is a 3D-tensor, through which spatial and temporal information is extracted.
- 3D-Max Pooling Layer: This layer is used to down-sample the input tensor along the specified axis.
- Batch-Normalization Layer: This is used to make the network faster and converge quickly and avoids internal covariate shift.
- Attention Layer: This layer helps to focus on the most informative aspect of input data during training. This allows the model to focus on what and where to look for the task at hand.
- Flatten Layer: This layer is used to take the input tensor and reshape it into a 1-D tensor which can be further given to the classifier for the classification task.

- Dense Layer: It is used to build a multi-layer perceptron which will be acting as the classifier for the proposed model.

The architecture is built using 5 3D-Convolutional Layers and 4 3D-Max Pooling layers of filter sizes as given in Figure 2. The model consists of 5 blocks for feature extraction. The classifier is built using 2 Fully Connected layers. Finally, the output layer is a sigmoid layer with a single neuron.

Inspired by the work of [9] we have used triplet attention in our model to extract richer and better feature representations. The attention module is computationally efficient in extracting channel and spatial attention. The module takes in the input tensor and returns a transformed tensor of a similar shape. The output of the module is element-wise multiplied by the input tensor to produce the final attention map.

ED-Net Model-B (CNN-LSTM based Model) Along with the 3D-CNN model, we also propose another CNN-LSTM-based model in order to compare the inference time between two different types of architecture. The proposed CNN-LSTM architecture is built using 5 2D-Convolutional layers with ReLU activation. For feature down-sampling, Max-Pooling is used with a filter size of 2. The dropout layer is used during the feature extraction process to introduce regularization. After the feature extraction, the tensor is flattened and to incorporate the sequential information LSTM layer is used before giving it to the classifier. The details of the model are given in Figure 3. To incorporate the temporal aspect, the TimeDistributed layer is used as provided in the Keras framework to apply the said layer to every temporal slice of the input. In this way, we apply the feature extraction using Conv2D on the temporal slices of N frames consecutively.

5 Experimental Setup

5.1 Implementation Details

The models are implemented using the Keras framework and TensorFlow backend. The data for the ”non-teaching” class is taken from different datasets including UCF-50 [12], Sports-1M [5], and Youtube-8M [1] to add to the diversity of videos. The clips from open-source places are taken that include classes like gaming, boxing, traveling, product review videos, football, cricket, etc to name a few. The training data contains no data imbalance between ”teaching” and ”non-teaching” class. The model is optimized using the Adam optimizer. The loss function used is binary-cross entropy as the task at hand is binary classification. The model is trained for 20 epochs and the results are presented in Section 6. The early stopping and reducing learning rate on the plateau are used. The model is trained on Nvidia Quadro RTX 6000, with a batch size of 16 only. The dataset is divided into an 8:1:1 ratio for training, validation, and testing. Apart from testing on the dataset, the model is also tested on random videos taken from open-source platforms to see the generalized capability of the model.

5.2 Evaluation Metrics

For evaluation and monitoring of our training and generalizability of the model, we have used Precision, Recall, and Accuracy as our collective metrics. The mathematical equation of each metric is given in the following equations. True Positive (TP) measures how many pixels the model is labeling to its correct class. False-positive (FP) measures how many pixels labeled to positive class belonged to the negative class. True Negative (TN) measures how many pixels classified as a negative class were actually from the negative class.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

6 Result and Discussion

The quantitative results are presented in Table 1. Model-A reaches the highest Accuracy of 98.87% while Model-B reaches an accuracy of 96.34%.Figure 4 and Figure 6 represents the loss curve for Model-A and Model-B. The graph shows that the model doesn't overfit or underfit. We have also tested our dataset using pre-trained models like ResNet-152 [4], Xception [3] and multi-head attention mechanism using implementation given at `https://keras.io/examples/vision/video_transformers/`. Figure 5 and Figure 7 represents the accuracy graph for Model-A and Model-B .For qualitative analysis, the results are checked using videos from the test set. In this, the prediction is performed on each individual frame of the input video. The results suggest possible failure cases of the models as well. The teaching videos generally have two characteristics, written texts, and a teacher or some annotation tool. The videos having texts or sentences in the videos which do not belong to the teaching category were misclassified as teaching videos. In our future work, we will try to fix this issue by including the audio information in feature extraction and classification tasks.

Table 1. Results on Teach-VID Dataset

Model	Precision	Recall	Accuracy
ResNet-152 + LSTM	96.09%	96.41%	97.77%
Xception + LSTM	99.15%	98.33%	99.53%
ViT	98.37%	98.37%	98.83%
ED-Net Model-A	99.22%	72.11%	98.87%
ED-Net Model-B	97.35%	92.45%	96.34%

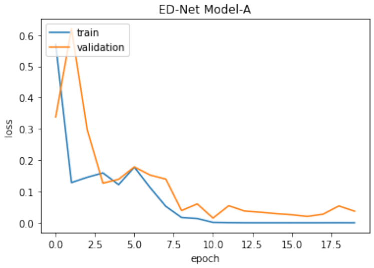

Fig. 4. Training and Validation Loss for ED-Net Model-A

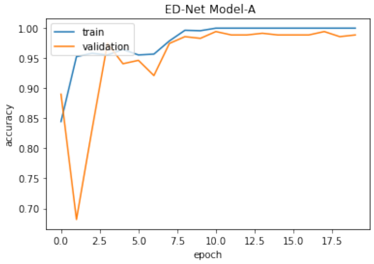

Fig. 5. Training and Validation Accuracy for ED-Net Model-A

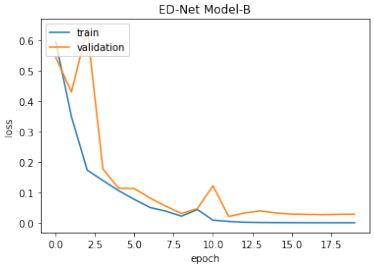

Fig. 6. Training and Validation Loss for ED-Net Model-B

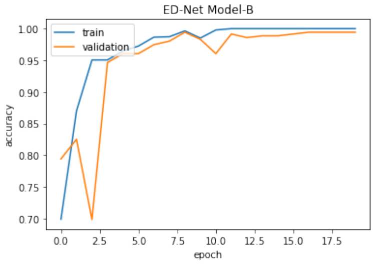

Fig. 7. Training and Validation Accuracy for ED-Net Model-B

7 Conclusion

In this paper, we have presented a novel dataset, Teach-VID, that contains video clips of 60 seconds each. The video clips are taken from an e-learning platform developed by our organization. This dataset is used to develop a system to classify Teaching and Non-Teaching videos. Along with the dataset, we have proposed two models for benchmarking purposes. The first model is developed using 3D-CNN and the second model is developed using 2D-CNN and LSTM.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Bhardwaj, S., Srinivasan, M., Khapra, M.M.: Efficient video classification using fewer frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 354–363 (2019)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
6. Li, X., Wang, Y., Zhou, Z., Qiao, Y.: Smallbignet: Integrating core and contextual views for video classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1092–1101 (2020)
7. Lin, R., Xiao, J., Fan, J.: Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
8. Long, X., De Melo, G., He, D., Li, F., Chi, Z., Wen, S., Gan, C.: Purely attention based local feature integration for video classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
9. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3139–3148 (2021)
10. Peng, Y., Zhao, Y., Zhang, J.: Two-stream collaborative learning with spatial-temporal attention for video classification. IEEE Transactions on Circuits and Systems for Video Technology **29**(3), 773–786 (2018)
11. Pouyanfar, S., Wang, T., Chen, S.C.: Residual attention-based fusion for video classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
12. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Machine vision and applications **24**(5), 971–981 (2013)

13. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5552–5561 (2019)
14. Wang, L., Li, W., Li, W., Van Gool, L.: Appearance-and-relation networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1430–1439 (2018)
15. Xu, X., Wu, X., Wang, G., Wang, H.: Violent video classification based on spatial-temporal cues using deep learning. In: 2018 11th International Symposium on Computational Intelligence and Design (ISCID). vol. 1, pp. 319–322. IEEE (2018)