1. Introduction

Based on the results of the ECCV workshop, we found that the two most popular paradigms for video classification are (i) *recurrent neural network* based methods and (ii) *cluster-and-aggregate* based methods. Not surprisingly, the third type of approaches based on C3D (3D convolutions) were not so popular because they are expensive in terms of their memory and compute requirements.

The author experimented with two different methods of training the *Teacher-Student* network.

- Serial Training
  The teacher is trained independently and then the student is trained to match the teacher with or without an appropriate regularizer to account for the classification loss.
- Parallel Training
  The teacher and student are trained jointly using classification loss as well as the matching loss.

The author experiment with different students

- a hierarchical RNN based model
- NetVLAD
- NeXtVLAD which is a memory efficient version of NetVLAD and was the best single model is the ECCV' 18 workshop.

2. Related Work

2.1 Video Classification

One of the popular datasets for video classification is the YouTube-8M dataset which

contains videos having an average length of 200 seconds.

- Long Short-Term Memory network (LSTM) is ued to encode the encode this sequence.

Explored methods:

1. feature aggregation in videos (temporal as well as spatial)
2. captureing the interactoins between labels
3. learning new non-linear units to model the interdependencies among activtions of the network

The state-of-the-art model on the 2017 version of the YouTube-8M dataset uses *NetVLAD* pooling to aggregate information from all the frames of a video.

2.2 Model Compression

We refer the reader to survey paper by “A Survey of Model Compression and Acceleration for Deep Neural Networks” for a thorough review of the field.

- Knowledge Distillation
- Quantization

3. Video Classification Models

The goal of video classification is to identify all the classes to which the video belongs.

3.1 Recurrent Network Based Models

- Hierarchical Recurrent Neural Network (H-RNN) based model which assumes that

each video contains a sequence of b equal sized blocks. Each of these blocks in ture is a sequence of l frames.
- The model contains a lower level RNN to encode each block (sequence of frames) and a higher level RNN to encode the video (sequence of blocks).

3.2 Cluster and Aggregate Based Models

- The author considered the *NetVLAD* model with Context Gating (CG) as proposed in “Learnable Pooling with Context Gating for Video Classification”.
    1. For every frame, it first assigns a **soft cluster** to the frame which results is a $M \times k$ dimensional representation. *A learnable clustering matrix* is introduced.
    2. This video representation is then fed to multiple fully connected layers with Context Gating (CG) which help to model interdependencies among network activations.

4. Proposed Approach

To design a simpler network *g* which looks at only a fraction of the *N* frames at inference time while still avoiding it to leverage the information from all the *N* frames at training time. To achieve this, a **Teacher-Student** network is proposed.

- **TEACHER**: The teacher network can be any state-of-the-art model described above (H-RNN, NetVLAD, NeXtVLAD). This teacher network looks at all the N frames of video and computes an encoding ξ_T of the video.
- **STUDENT**: A student network which only processes every j-th frame of the video and computes a representation ξ_S.
    1. Both the teacher and the student have the same architecture; the parameters of the output layer are shared between the teacher and the student.
    2. They also tried a simple variant of the model, where in addition to ensuring that the final representations ξ_T and ξ_S are similar, they also ensure that the intermediate representations (I_T and I_S) of the models are similar.

- **TRAINING**: *Serial* mode and *Parallel* mode

5. Experimental Setup

1. **Dataset**: YouTube-8M dataset (2017 version)
2. **Hyperparameters**:
    1. Adam Optimizer with the initial learning rate set to 0.001 and then decrease it exponentially with 0.95 decay rate.
    2. Batch size
    3. A 2-layered MultiLSTM Cell
    4. dropout=0.5 and L_2 regularization
    5. Train for 5 epochs.
3. **Evaluation Metrics**:
    1. GAP (Global Average Precision)
    2. mAP (Mean Average Precision)
4. **Model Compared**:
   The authors compared their **Teacher-Student** network with the following models:
    - Teacher-Skyline
    - Baseline Methods

6. Discussion and Results

7. Conclusion and Future Work

- Proposed a method to reduce the computation time for video classification using the idea of **distillation**.
- They first train a teacher network which computes a representation of the video

using all the frames in the video, then a student network which only processes k frames of the video.

- They use different combinations of loss functions which ensures that
  - the final representation produced by the student network is similar as that produced by the teacher
  - the output probability distributions produced by the student are similar to those produced by the teacher
- They compare the proposed models with a strong baseline and skyline and the proposed models outperforms the baseline and skyline models.