CVPR 2019: Supplementary Material

Paper ID 6940

1 Architecture of *Teacher-Student* Framework in *NeXtVLAD* Model :

As shown in Figure 1, we have trained the student network (using $k{=}\frac{N}{j} - 1$ frames) from the pretrained teacher network of *NeXtVLAD* (which uses all the N frames in a video). In the training of this framework, $\mathscr{L}_{rep}$ denotes the representation loss between the concatenated video encodings $\mathcal{E}_T$ and $\mathcal{E}_S$ from teacher and student network respectively. To take full benefit of knowledge-distillation, $\mathscr{L}_{pred}$ (KL divergence between output probability distributions from teacher and student) and $\mathscr{L}_{CE}$ (standard classification loss) are optimized simultaneously while training the student network. The results of this experiment have been reported in the main paper.

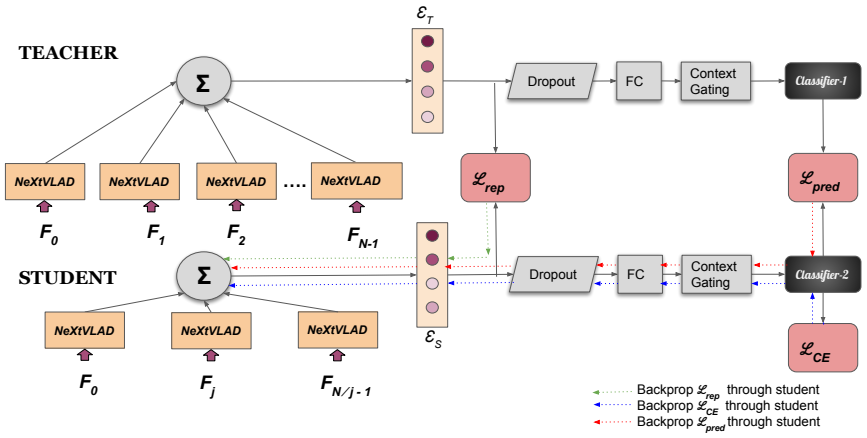

Figure 1: Architecture of Teacher-Student framework in *NeXtVLAD* model

2 Performance of *Teacher-Student* Framework on the Ensemble of Models:

As mentioned in *NeXtVLAD* paper, for an ensemble training of *Teacher-Student* architecture, we have used a batch size of 160 with an ensemble of 3 models. Rest of the hyperparameter setting is same as in a single model experimentation. We train the ensemble of student networks (3 networks) with distillation loss $\mathscr{L}_{pred}$ from the weighted soft-targets from on-the-fly ensemble of three teachers. Alongwith the $\mathscr{L}_{pred}$, each of the student networks is trained with $\mathscr{L}_{rep}$ (representation learning) loss in order to mimic the video representations from their corresponding teachers. As we can see in table 1, the use of distillation is beneficial even in the case of ensemble methods. We can still explore more hyperparameter settings for better performance of the *Student*.

Model: NeXtVLAD	mAP	GAP	Steps
Ensemble-Teacher	0.485	0.839	310k
Ensemble-Uniform	0.445	0.821	384k
Ensemble-Student	0.450	0.825	306k

Table 1: Training of ensemble models with k=30 frames

3 Analysis of *Teacher-Student* Framework on *Rare* and *Frequent* Classes in the Dataset:

To do further analysis of *Student* with $Uniform$ baseline, we examined the performance of these networks (RNN-based models) on bottom-r (rare) and top-r (frequent) classes according to the available training data. As reported in table 2, the performance of model largely depends on the most frequent classes and there is an approximate gap of 0.5-1% between the *Student* and $Uniform$. On the other hand in case of rare (less-frequent) classes, the *Student* network clearly beats the baseline by a slightly bigger margin. This strengthens our intuition that even while working with fewer frames, the *Student* network manages to perform well on the rare classes in the dataset with the help of *Knowledge Distillation*.

bottom-#r classes		Model			top-#r classes		Model		
#RareClasses	Performance	Teacher	Uniform	Student	#FrequentClasses	Performance	Teacher	Uniform	Student
bottom-100	mAP	0.005	0.005	0.005	top-100	mAP	0.018	0.018	0.018
	GAP	0.293	0.249	0.279		GAP	0.918	0.909	0.915
bottom-200	mAP	0.013	0.011	0.013	top-200	mAP	0.035	0.034	0.035
	GAP	0.300	0.247	0.281		GAP	0.901	0.892	0.899
bottom-500	mAP	0.033	0.028	0.031	top-500	mAP	0.077	0.075	0.077
	GAP	0.302	0.248	0.280		GAP	0.873	0.862	0.870
bottom-1000	mAP	0.065	0.056	0.062	top-1000	mAP	0.137	0.132	0.135
	GAP	0.298	0.248	0.282		GAP	0.852	0.839	0.848
bottom-2000	mAP	0.132	0.116	0.128	top-2000	mAP	0.232	0.220	0.229
	GAP	0.313	0.268	0.297		GAP	0.832	0.817	0.828

Table 2: Comparison of Student network with Uniform-k baseline with k=30 frames on #r most rare (bottom) and most frequent (top) classes in the dataset.