Netid: vdongre2

IE 534 / CS 547
Assignment # 9
Video Recognition: Human Action Recognition

Part 1: Single Frame Model
(**code:** singleframe_train.py, singleframe_test.py, helperFunctions.py)

Part 2: 3d Resnet Model
(**code:** 3d_resnet_train.py, 3d_resnet_test.py, helperFunctions.py)

Part 3: Combined Model (Ensemble/Average)
(**code:** combined_performance.py)

Q.1 : There are three sets of results for comparison: 1.) single-frame model, 2.) 3D model, 3.) combined output of the two models. For each of these three, report the (top1_accuracy,top5_accuracy,top10_accuracy): Did the results improve after combining the outputs?

Single Frame Model:
Training Accuracy: 96.77011494252874 (10th epoch)
Test Accuracy: 73.48648648648648 (10th epoch)

Sequence Model (3d) : (Model trained for 6 epochs as it starts converging after that)
Training Accuracy: 98.88888888888889 (6th epoch)
Test Accuracy: 83.84533898305085 (6th epoch)

Performance

Model\Parameter	Top 1 accuracy	Top 5 accuracy	Top 10 accuracy
Single Frame Model	0.786413	0.947925	0.976738
3d Model	0.808971	0.954391	0.982147
Combined Model	0.857256	0.977002	0.988633

We can see that the combined model performs better than the individual methods. So, we can say that the results do improve after combining the outputs.

Q.2: Use the confusion matrices to get the 10 classes with the highest performance and the 10 classes with the lowest performance: Are there differences/similarities? Can anything be said about whether particular action classes are discriminated more by spatial information versus temporal information?

Model	Highest Performing	Lowest Performing
Single Frame Model	BabyCrawling 1.0 HorseRace 1.0 Surfing 1.0 BasketballDunk 1.0 Skijet 1.0 Billiards 1.0 Rowing 1.0 RockClimbingIndoor 1.0 PoleVault 1.0 PlayingFlute 1.0	JumpRope 0.02631579 BodyWeightSquats 0.1 HandstandWalking 0.20588236 FrontCrawl 0.24324325 YoYo 0.30555555 HighJump 0.35135135 CricketShot 0.36734694 Nunchucks 0.4 BrushingTeeth 0.44444445 Hammering 0.4848485
Sequence Model (3d)	WritingOnBoard 1.0 45.0 Diving 1.0 45.0 Drumming 1.0 45.0 VolleyballSpiking 1.0 35.0 UnevenBars 1.0 28.0 Fencing 1.0 34.0 TrampolineJumping 1.0 32.0 PlayingViolin 1.0 28.0 JumpingJack 1.0 37.0 PlayingTabla 1.0 31.0	HighJump 0.2972973 37.0 Lunges 0.3243243 37.0 Nunchucks 0.4 35.0 SoccerJuggling 0.41025642 39.0 Haircut 0.42424244 33.0 CricketShot 0.42857143 49.0 FrontCrawl 0.45945945 37.0 BrushingTeeth 0.4722222 36.0 CricketBowling 0.4722222 36.0 PizzaTossing 0.4848485 33.0
Combined Model	Skijet 1.0 PlayingPiano 1.0 Billiards 1.0 PlayingTabla 1.0 PlayingViolin 1.0 Rowing 1.0 PoleVault 1.0 RockClimbingIndoor 1.0 Bowling 1.0 HorseRace 1.0	HighJump 0.3243243 Nunchucks 0.37142858 BrushingTeeth 0.41666666 FrontCrawl 0.43243244 Hammering 0.4848485 CricketShot 0.48979592 YoYo 0.5 CricketBowling 0.5 Lunges 0.5135135 PizzaTossing 0.5151515

We can see some of the activities that have nearly similar probabilities assigned by both models and are not amongst the High/Worst performing classes, need more spatial and temporal information. In some cases, it is obvious that temporal information is more important than the spatial information as these activities are better recognized by the 3d model than the single frame model as the 3d model assigns a higher probability to such activities. For example, JumpRope 0.02631579 in Single Frame Model, JumpRope 0.9736842 in Sequence Model shows more need of temporal information.

In cases where single frame model is performing better than the 3d model more spatial data is needed. For example: SoccerJuggling 0.6666667 in Single Frame Model and SoccerJuggling 0.41025642 in Sequence Model shows there is need of more spatial data.

Overall the performance of combined model is better than both models. One interesting observation is that activities in which something is being “played” are almost correctly identified by all the models.

Q3. Use the confusion matrices to get the 10 most confused classes. That is, which off-diagonal elements of the confusion matrix are the largest: Are there any notable examples? (code: combined_performance.py)

Model	Most Confused Classes
Single Frame Model	('FrontCrawl', 'BreastStroke'), ('BrushingTeeth', 'ShavingBeard'), ('CricketShot', 'CricketBowling'), ('Hammering', 'HeadMassage'), ('BodyWeightSquats', 'Lunges'), ('BodyWeightSquats', 'WallPushups'), ('JumpRope', 'HulaHoop'), ('HammerThrow', 'ThrowDiscus'), ('BalanceBeam', 'ParallelBars'), ('JavelinThrow', 'LongJump')
Sequence Model (3d)	('Haircut', 'BlowDryHair'), ('FrontCrawl', 'BreastStroke'), ('BrushingTeeth', 'ShavingBeard'), ('Nunchucks', 'TaiChi'), ('BoxingPunchingBag', 'BoxingSpeedBag'), ('PommelHorse', 'ParallelBars'), ('YoYo', 'JugglingBalls'), ('HighJump', 'LongJump'), ('HammerThrow', 'ThrowDiscus'), ('PlayingCello', 'PlayingViolin')
Combined Model	('FrontCrawl', 'BreastStroke'), ('BrushingTeeth', 'ShavingBeard'), ('Haircut', 'BlowDryHair'), ('PommelHorse', 'ParallelBars'), ('Nunchucks', 'TaiChi'), ('CricketShot', 'CricketBowling'), ('YoYo', 'JugglingBalls'), ('HammerThrow', 'ThrowDiscus'), ('BoxingPunchingBag', 'BoxingSpeedBag'), ('Hammering', 'HeadMassage')

We see some common confusing classes among the three models: ('HammerThrow', 'ThrowDiscus'); ('BrushingTeeth', 'ShavingBeard'); ('FrontCrawl', 'BreastStroke')