# CS 547 Deep Learning

# Homework 9 Video Recognition Report

Name: Jiashuo Tong

Date Submitted: Dec. 15, 2019

**Question #1:** Did the results improve after combining the outputs?

Table 1. Summary for Performances of Different Models

| Model | Time for Testing | (Top-1 accuracy, Top-5 accuracy, Top-10 accuracy) |
|---|---|---|
| Single Frame Model | 7196 s | (0.785091,0.946603,0.974888) |
| Sequence Model | 109617 s | (0.825535,0.961671,0.979117) |
| Combined Model | - | (0.830029,0.969072,0.983347) |

**Note:** The combined model does not need to be tested on the test set. We can simply take the average of the confusion matrices for the Single Frame Model and the Sequence Model to obtain the confusion matrix for the Combined Model.

According to Table 1, the combined model has a slightly better performance. The Top-1 accuracy for the combined model is the highest among the three, reaching a value of 0.830029. The Top-5 accuracy and the top-10 accuracy are also improved.

**Question #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?

Table 2: Ten Categories with Lowest Performance and Highest Performance

| **Single Frame Model** | | **Sequence Frame Model** | | **Combined Model** | |
|---|---|---|---|---|---|
| **10 Lowest** | **10 Highest** | **10 Lowest** | **10 Highest** | **10 Lowest** | **10 Highest** |
| HandstandWalking<br>0.029411765 | Skijet<br>1.0 | YoYo<br>0.11111111 | SumoWrestling<br>1.0 | JumpRope<br>0.10811525 | SumoWrestling<br>1.0 |
| JumpRope<br>0.078947365 | HorseRace<br>1.0 | HammerThrow<br>0.33333334 | TrampolineJumping<br>1.0 | HandstandWalking<br>0.23050489 | TrampolineJumping<br>1.0 |
| BodyWeightSquats<br>0.16666667 | TrampolineJumping<br>1.0 | HighJump<br>0.3783784 | WallPushups<br>1.0 | YoYo<br>0.29817605 | WallPushups<br>1.0 |
| YoYo<br>0.19444445 | BabyCrawling<br>1.0 | Shotput<br>0.39130434 | Billiards<br>1.0 | BodyWeightSquats<br>0.37774253 | Billiards<br>1.0 |
| HighJump<br>0.27027026 | BasketballDunk 1.0 | Nunchucks<br>0.4 | BabyCrawling<br>1.0 | HighJump<br>0.39994713 | BabyCrawling<br>1.0 |
| FrontCrawl 0.2972973 | Diving<br>1.0 | PizzaTossing<br>0.42424244 | Skijet<br>1.0 | PizzaTossing<br>0.43431139 | Skijet<br>1.0 |
| Nunchucks 0.34285715 | PoleVault<br>1.0 | Hammering<br>0.42424244 | PoleVault<br>1.0 | Nunchucks<br>0.45043616 | PoleVault<br>1.0 |
| CricketShot<br>0.42857143 | BreastStroke<br>1.0 | Lunges<br>0.43243244 | HorseRace<br>1.0 | CricketShot<br>0.46021676 | HorseRace<br>1.0 |
| PullUps<br>0.42857143 | SumoWrestling<br>1.0 | LongJump<br>0.43589744 | BasketballDunk<br>1.0 | Lunges<br>0.47343378 | BasketballDunk<br>1.0 |
| JumpingJack<br>0.45945945 | Billiards<br>1.0 | CricketBowling<br>0.4722222 | Diving<br>1.0 | CricketBowling<br>0.49537404 | Diving<br>1.0 |

**Note:** For the “10 Highest” categories, there are more than 10 categories with perfect accuracies for both Single

Frame Model and Sequence Model. Therefore, the categories that are common for both models are picked, while most of the categories that only appear once are deleted.

As Table 2 shows, the poorest performance categories that are common for both models are Yoyo, HighJump, and Nunchucks. The best performance categories that are common for both models are Skijet, Horcerace, TrampolineJumping, BabyCrawling, BasketballDunk, Diving, PoleVault, SumoWrestling, and Billiards.

The best performance categories share one similarity, that is, most of the human actions must be performed in a certain background or with certain objects. For example, “Diving” should always be done in sea water; “HorseRace” cannot be done without a horse; “BasketballDunk” is impossible without a basketball hoop; “Billiards” must come with billiard boards. Therefore, the background or the objects should be the patterns that are easily identified by deep learning models.

The poorest performance categories, in contrast, are often the kinds of actions that people would like to do in many places. For example, people can “HighJump”, “Yoyo”, or “Handstand walk” in homes, in gyms, and on the playgrounds. In such cases, the deep learning model have no constant item to spot, so it has to learn the action itself, which is a much more difficult task.

Comparing the accuracies for “HighJump” and “Nunchucks” from the two “10 Lowest” columns, we find that the sequence model can significantly improve the accuracy. Comparing the other human actions, we find that the sequence model shows a reduced performance on certain actions. For example, the Sequence Model achieves an accuracy of 0.424242 for “PizzaTossing” and “Hammering”, while the Single Frames Model achieves well above 0.542857 and 0.515151, respectively. Maybe the reason is that the Sequence Model is poor at identifying the pizzas in “PizzaTossing” or the hammers in “Hammering”. Therefore, the actions that highly rely on a certain tool are usually discriminated by spatial versus temporal information.

Overall, the Combined Model achieve a better accuracy than both the Single Frames Model and the Sequence Model.

**Question #3:** 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?

The following table is retrieved from the confusion matrices of the three models.

Table 3: Human Action Categories That Are Most Mistaken for Another Category

| **Single Frame Model** | | | **Sequence Model** | | | **Combined Model** | | |
|---|---|---|---|---|---|---|---|---|
| **Category** | **Mistaken for** | **Probabilty** | **Category** | **Mistaken for** | **Probability** | **Category** | **Mistaken for** | **Probability** |
| BreastStroke | FrontCrawl | 0.432432 | BreastStroke | FrontCrawl | 0.702703 | BreastStroke | FrontCrawl | 0.459459 |
| BrushingTeeth | ApplyLipstick | 0.40625 | BlowDryHair | Haircut | 0.575758 | ParallelBars | PommelHorse | 0.4 |
| ParallelBars | BalanceBeam | 0.387097 | ApplyLipstick | ShavingBeard | 0.372093 | BrushingTeeth | ApplyLipstick | 0.375 |
| ParallelBars | PommelHorse | 0.371429 | ParallelBars | PommelHorse | 0.371429 | BlowDryHair | Haircut | 0.363636 |
| BlowDryHair | Haircut | 0.30303 | JavelinThrow | HighJump | 0.351351 | ParallelBars | BalanceBeam | 0.354839 |
| HulaHoop | JumpRope | 0.289474 | ApplyLipstick | BrushingTeeth | 0.333333 | HeadMassage | Hammering | 0.30303 |
| CricketBowling | CricketShot | 0.285714 | CricketBowling | CricketShot | 0.285714 | HulaHoop | JumpRope | 0.289474 |
| HeadMassage | Hammering | 0.272727 | SoccerJuggling | Swing | 0.285714 | CricketBowling | CricketShot | 0.285714 |
| JavelinThrow | HighJump | 0.27027 | Skijet | Rowing | 0.277778 | JavelinThrow | HighJump | 0.27027 |
| ApplyLipstick | ApplyEyeMakeup | 0.25 | BoxingSpeedBag | BoxingPunchingBag | 0.265306 | ThrowDiscus | HammerThrow | 0.244444 |

**Note:** The category pairs that are highlighted in yellow are the confused classes that are common for all the three models.

According to Table 3, we notice that there are 6 common confused pairs for all the three models. They are (BreastStroke, FrontCrawl), (BrushingTeeth, ApplyLipstick), (ParallelBars, PommelHorse), (BlowDryHair, Haircut), (CricketBowling, CricketShot), and (JavelinThrow, HighJump).

We have the observation that the two actions for a certain confused pair usually involve the same subjects. For example, both BlowDryHair and Haircut have hair as the subject; BrushingTeeth and ApplyLipstick both need to be conducted on a face; BreastStroke and FrontCrawl are both swimming styles.