Derin Öğrenme Yöntemleriyle Video Görüntülerden Aksiyon Tanıma

Recep Furkan Koçyiğit

22501048

furkan.kocyigit@std.yildiz.edu.tr

Bilgisayar Mühendisliği Bölümü

Elektrik Elektronik Fakültesi, Yıldız Teknik Üniversitesi

Özet

Bu ödevde, içerisinde masa tenisi şutu, tenis raket sallama, yumruk atma, rafting ve sörf, kategorilerinde toplamda 2111 adet video içeren UCF-101 veri kümesinde Konvolüsyonel Sinir Ağları (3D CNN) ve Uzun-Kısa Vadeli Bellek (LSTM) ile sınıflandırma işlemi yapılmıştır. Ödev Python programlama dili kullanılarak yazılmıştır. Videodaki görüntüler eğitim süresi göz önüne alınarak 56 genişlik, 56 yüksekliğe boyutlandırılmıştır. Her iki modelin de yüksek başarı sağladığı görülmüştür. LSTM modellerinin yapısı gereği eğitim ve test işlemlerinin daha fazla zaman aldığı gözlenmiştir.

Giriş

Ödev kapsamında aşağıdaki hiper parametreler arasından en iyi başarıyı sağlayan model bulunmaya çalışılmıştır. Eğitim için 1215, validasyon için 304 ve test için 592 adet veri kullanılmıştır. Veriye bakıldığında düzgün bir dağılım gösterdiği için doğruluk (accuracy) kullanılmıştır.

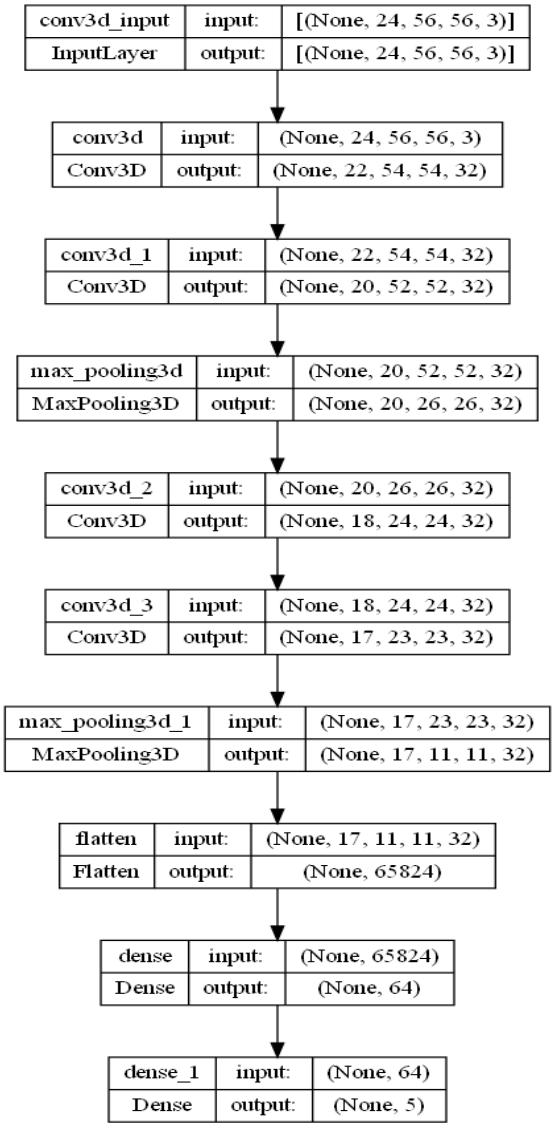
1. Görüntü özellikleri: 56 genişlik, 56 yükseklik, 3 kanal
2. Konvolüsyon Katmanı = [4]
3. Filtre Sayısı = [32]
4. Kernel Boyutu = [3x3
5. Dropout = [Var, Yok]
6. Mini Batch Boyutu = 16
7. Epoch = 5
8. Konvolüsyon Katmanı Aktivasyon Fonksiyonu = [ReLU, ELU, Sigmoid, Tanh]
9. Optimizasyon Algoritması = Adam
10. Çerçeve sayısı (Frame) = [12, 24, 36, 48]

Ödev kapsamında yapılan iş parçacıkları şu şekildedir:

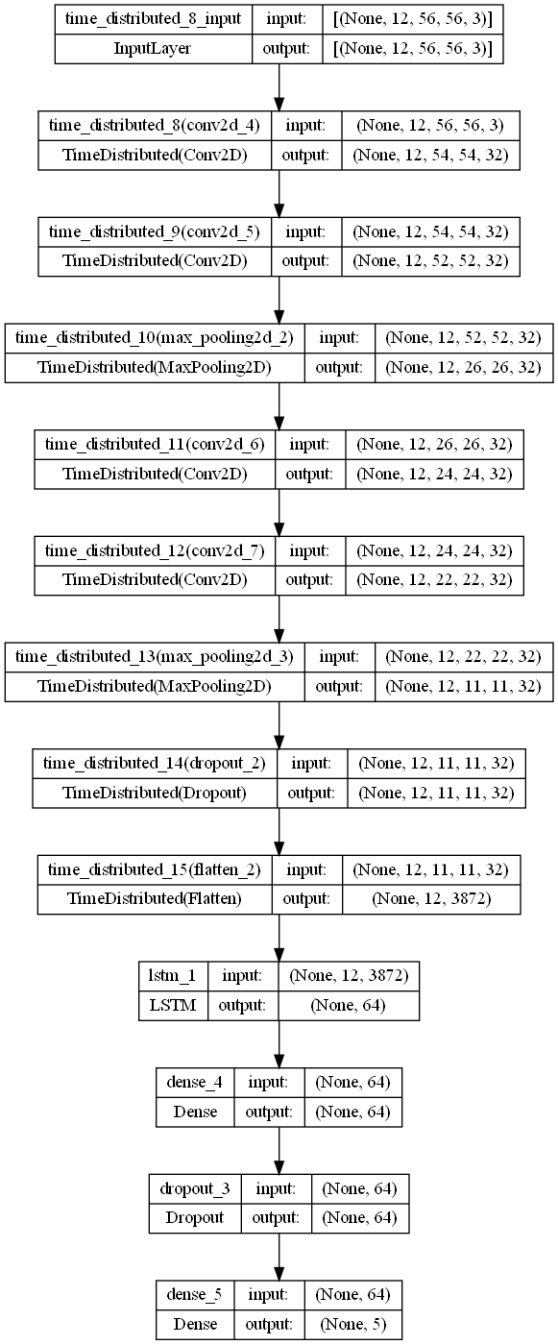
1. **Verinin Yüklenmesi:** Video isimleri bulunan eğitim ve test dosyalarından bu isimlerin ve kategorilerinin ödev kapsamındaki kategorilerin filtrelenerek alınması.
2. **Her Kategori İçin Ortalama Çerçeve Sayısının Bulunması:** Dosya ismi verilen videolar için kategori bazında ortalama çerçeve sayılarının bulunması.
3. **Verinin Ön İşleme Aşamasının Gerçeklenmesi:** Verilen video dosyasının ismi için dosyanın belirli bir çerçevede okunması ve 0-255 arasındaki bit değerlerinin normalize edilmesi, çıktılara One-Hot Encode yöntemi uygulanması ve eğitim, validasyon ve test olarak ayrılması.
4. **Çerçeve Sayısı, Aktivasyon Fonksiyonu ve Dropout Bilgisine Göre Modellerin Oluşturulması:** 3D CNN ve CNN+LSTM ağlarını parametrik olarak oluşturmak için gerekli fonksiyonların oluşturulması.
5. **Model Eğitiminin Görselleştirilmesi:** Eğitilen modelin başarı ve kayıplarının eğitim ve validasyon için gösterilmesi
6. **Sonuçların yazdırılması:** Eğitimi tamamlanan modellerin karmaşıklık matrislerinin gösterilmesi ve doğruluk(accuracy), f1 skoru, duyarlılık(recall) ve kesinlik(precision) bilgilerini göstermek için yazılmıştır.
7. **Hiper Parametrelere Göre Modellerin Oluşturulması Sonuçların Elde Edilmesi:** Verilen hiper parametrelere için tüm kombinasyonlarına göre modeller oluşturulması.
8. **En İyi Modelin Eğitim Kümesi ve Validasyon kümesi ile Eğitilip Başarısının Bulunması:** En iyi başarı oranına sahip modelin hiper parametreleri belirlendikten sonra model oluşturulur ve eğitim ve validasyon kümesinin birleşimi ile tekrar eğitilir. Daha sonra başarısı hesaplanır.

Yöntem

En başarılı 3D CNN ağı 4 konvolüsyon katmanı, 32 adet filtre içermekte, kernel boyutu 3x3, aktivasyon fonksiyonu ReLU ve dropout içermeyen, 2 adet Max Pooling katmanı içeren ve 24 çerçeveye sahip görüntülerle eğitilen model olmuştur. Modele Max Pooling işlemi uygulandığında başarının arttığı ve parametre sayısının azalmasından dolayı işlem maliyetinin de azaldığı görülmüştür. Modelin blok diyagramı aşağıdaki gibidir:

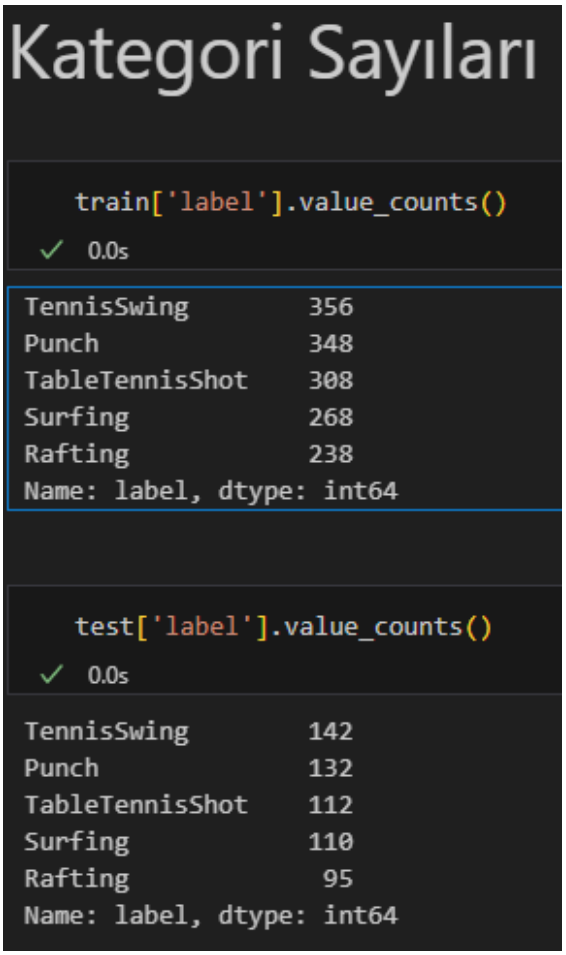


En başarılı CNN+LSTM ağı 4 konvolüsyon katmanı, 32 adet filtre içermekte, kernel boyutu 3x3, aktivasyon fonksiyonu Tanh ve dropout içeren, 2 adet Max Pooling katmanı içeren ve 12 çerçeveye sahip görüntülerle eğitilen model olmuştur. Modelin blok diyagramı aşağıdaki gibidir:

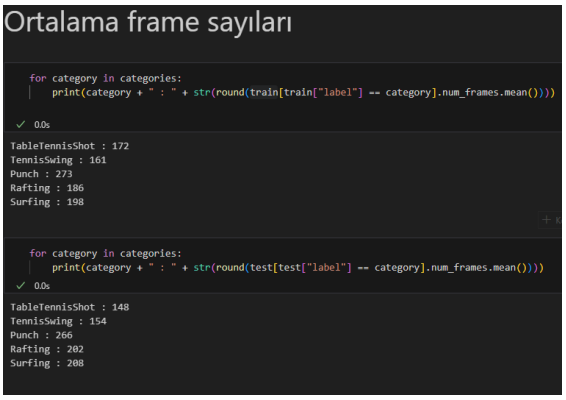


Uygulama

UCF-101 veri kümesinde istenilen 5 kategori için toplamda 1519 adet eğitim, 592 adeti test verisidir. Eğitim ve test kümesindeki kategori dağılım aşağıda verilmiştir:



Eğitim ve test verilerinde her kategori için ortalama çerçeve sayıları aşağıda verilmiştir:



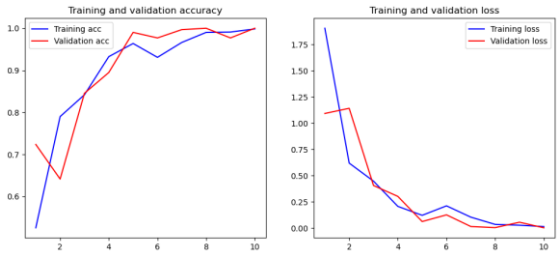
Validasyon verisi ile yapılan testler sonucunda modellerin başarısı aşağıda verilmiştir:

Model No	Model	Frame	Aktivasyon	Dropout	Accuracy	F1
62	lstm	12	tanh	True	0.9835526315789473	0.9803850509984156
35	cnn	24	relu	False	0.9802631578947368	0.9776994198265866
3	cnn	48	relu	False	0.9802631578947368	0.9766886828530641
64	lstm	12	tanh	False	0.9736842105263158	0.9711488405374986
50	lstm	12	relu	True	0.9703947368421053	0.9683217049176127
55	cnn	12	elu	False	0.9703947368421053	0.9662735516291516
34	lstm	24	relu	True	0.9638157894736842	0.961184922717639
53	cnn	12	elu	True	0.9671052631578947	0.960410520384516
40	lstm	24	elu	False	0.9671052631578947	0.9592647658208444
54	lstm	12	elu	True	0.9605263157894737	0.9588475690931499
30	lstm	36	tanh	True	0.9572368421052632	0.9507596756432557
17	cnn	36	relu	True	0.9539473684210527	0.950163342445412
46	lstm	24	tanh	True	0.9539473684210527	0.9491356561644082
37	cnn	24	elu	True	0.9572368421052632	0.9472470183538146
52	lstm	12	relu	False	0.9572368421052632	0.9468986921816477
56	lstm	12	elu	False	0.9342105263157895	0.9286181821255065
19	cnn	36	relu	False	0.930921052631579	0.9265103575343095
2	lstm	48	relu	True	0.9342105263157895	0.9224393576376952
38	lstm	24	elu	True	0.930921052631579	0.922187354979117
49	cnn	12	relu	True	0.9210526315789473	0.9157593467930486
18	lstm	36	relu	True	0.9243421052631579	0.9111830723200199
20	lstm	36	relu	False	0.9177631578947368	0.907969168294528
48	lstm	24	tanh	False	0.9078947368421053	0.9006780454203671
23	cnn	36	elu	False	0.9078947368421053	0.8983173753159728
14	lstm	48	tanh	True	0.9013157894736842	0.8929164776361216
33	cnn	24	relu	True	0.8881578947368421	0.880151004251553
8	lstm	48	elu	False	0.881578947368421	0.8715321939210773
1	cnn	48	relu	True	0.819078947368421	0.8216330091601594
32	lstm	36	tanh	False	0.8289473684210527	0.8202816061453511
22	lstm	36	elu	True	0.8223684210526315	0.814199309840259
51	cnn	12	relu	False	0.8026315789473685	0.8058901458159189
6	lstm	48	elu	True	0.8125	0.7919915914045468
4	lstm	48	relu	False	0.7861842105263158	0.7557002116043974
36	lstm	24	relu	False	0.7730263157894737	0.7544878073549027
24	lstm	36	elu	False	0.6644736842105263	0.6074366651312759
21	cnn	36	elu	True	0.5822368421052632	0.5012607152074858
5	cnn	48	elu	True	0.2532894736842105	0.08083989501312336
7	cnn	48	elu	False	0.2532894736842105	0.08083989501312336
9	cnn	48	sigmoid	True	0.2532894736842105	0.08083989501312336
10	lstm	48	sigmoid	True	0.2532894736842105	0.08083989501312336
11	cnn	48	sigmoid	False	0.2532894736842105	0.08083989501312336
12	lstm	48	sigmoid	False	0.2532894736842105	0.08083989501312336
13	cnn	48	tanh	True	0.2532894736842105	0.08083989501312336
25	cnn	36	sigmoid	True	0.2532894736842105	0.08083989501312336
26	lstm	36	sigmoid	True	0.2532894736842105	0.08083989501312336
27	cnn	36	sigmoid	False	0.2532894736842105	0.08083989501312336
29	cnn	36	tanh	True	0.2532894736842105	0.08083989501312336
41	cnn	24	sigmoid	True	0.2532894736842105	0.08083989501312336
43	cnn	24	sigmoid	False	0.2532894736842105	0.08083989501312336
44	lstm	24	sigmoid	False	0.2532894736842105	0.08083989501312336
57	cnn	12	sigmoid	True	0.2532894736842105	0.08083989501312336
58	lstm	12	sigmoid	True	0.2532894736842105	0.08083989501312336
59	cnn	12	sigmoid	False	0.2532894736842105	0.08083989501312336
61	cnn	12	tanh	True	0.2532894736842105	0.08083989501312336
15	cnn	48	tanh	False	0.24342105263157895	0.0783068783068783
16	lstm	48	tanh	False	0.24342105263157895	0.0783068783068783
28	lstm	36	sigmoid	False	0.24342105263157895	0.0783068783068783
31	cnn	36	tanh	False	0.24342105263157895	0.0783068783068783
39	cnn	24	elu	False	0.24342105263157895	0.0783068783068783
42	lstm	24	sigmoid	True	0.24342105263157895	0.0783068783068783
45	cnn	24	tanh	True	0.24342105263157895	0.0783068783068783
47	cnn	24	tanh	False	0.24342105263157895	0.0783068783068783
60	lstm	12	sigmoid	False	0.24342105263157895	0.0783068783068783
63	cnn	12	tanh	False	0.24342105263157895	0.0783068783068783

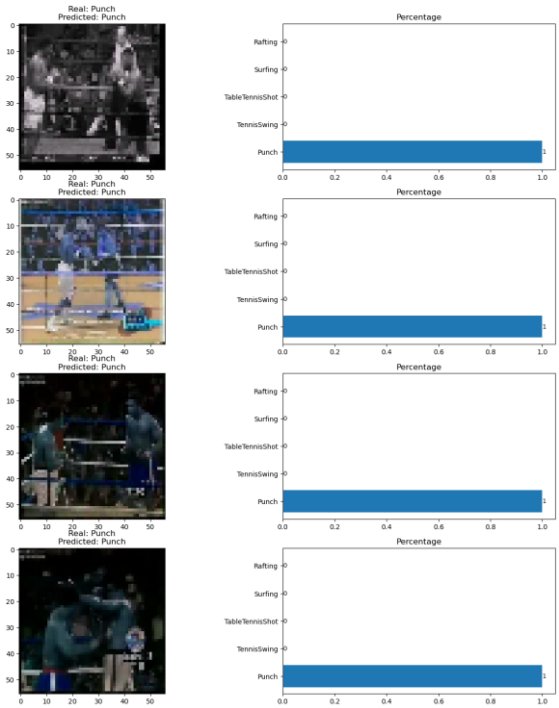
Tabloya bakıldığında aktivasyon fonksiyonu Sigmoid ve Tanh olan modellerin öğrenme konusunda diğer aktivasyon fonksiyonlarından başarısız olduğu görülmektedir. Ayrıca 48 çerçeve sayısına sahip görüntülerle eğitilen modellerin genel olarak daha düşük başarıya sahip olduğu görülmüştür. Tabloya bakarak dropout bilgisinin başarıda bir artışa ya da kayba sebep olduğu söylenemez. Aynı şekilde 3D CNN ve CNN+LSTM modellerinin de başarısı için kesin olarak birbirlerine üstünlüğü olduğu söylenemez.

3D CNN Modeli

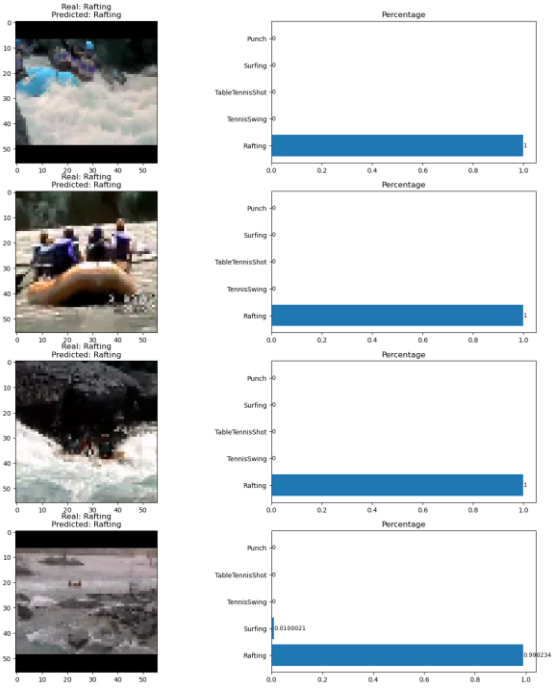
En iyi parametrelere sahip 3D CNN ağının eğitim grafiği aşağıdaki gibidir:



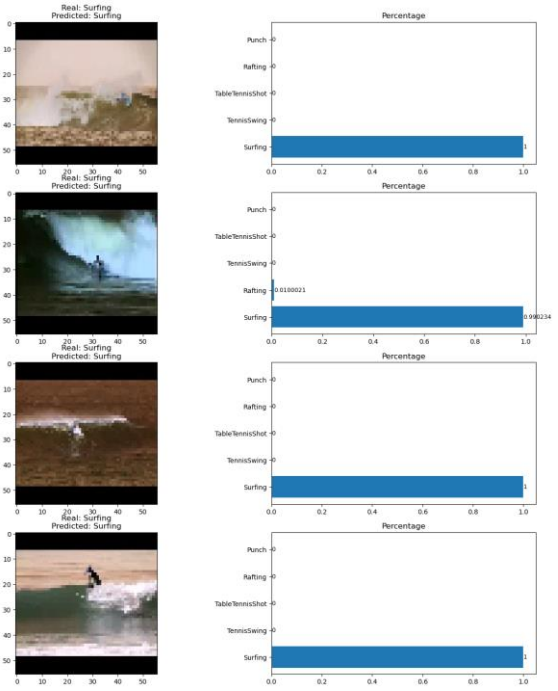
Burada eğitim ve validasyon başarılarının 1' e geldiği görülmektedir. Eğitim süresi 230,78 saniye sürmüştür. Test verisiyle test edildiğine başarı oranı %100 bulunmuştur. Punch kategorisine ait 4 farklı rastgele örnekle elde edilen sonuçlar şu şekildedir:



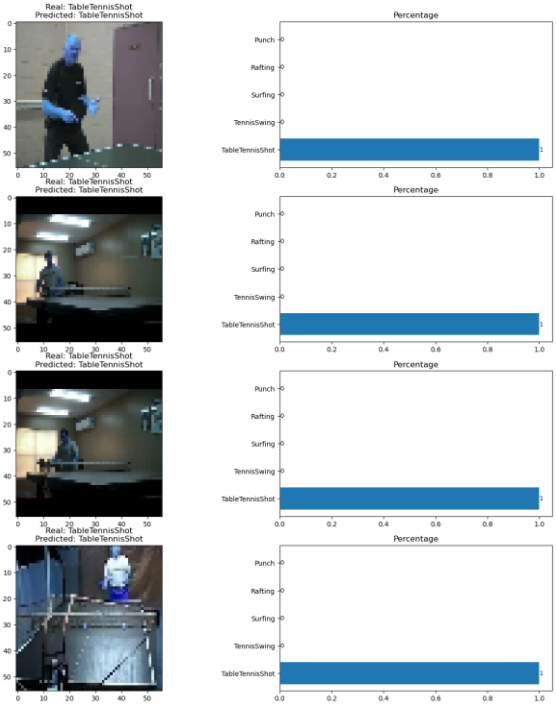
4 farklı örneğe baktığımızda da modelin kesin olarak doğru bildiği görülmektedir. Rafting kategorisine ait 4 farklı rastgele seçilen örneğe ait sonuçlar şu şekildedir:



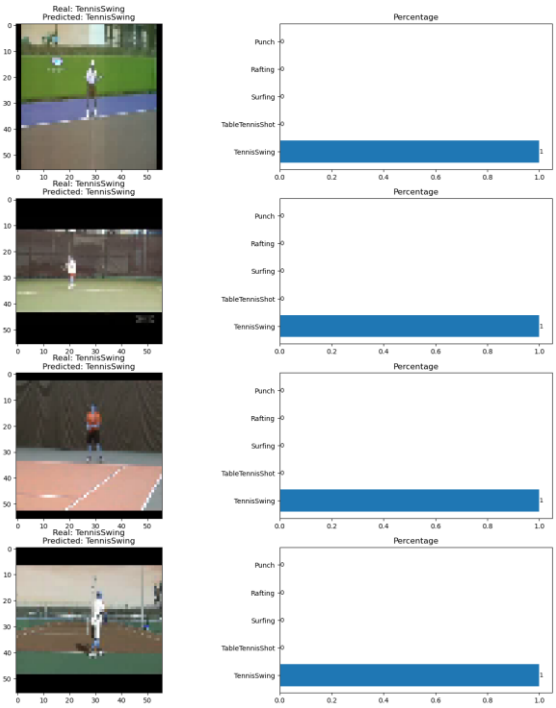
Burada da modelin raftingi kesin olarak bildiği görülmektedir. Sörf kategorisine ait 4 farklı rastgele seçilen resim aşağıdaki gibidir:



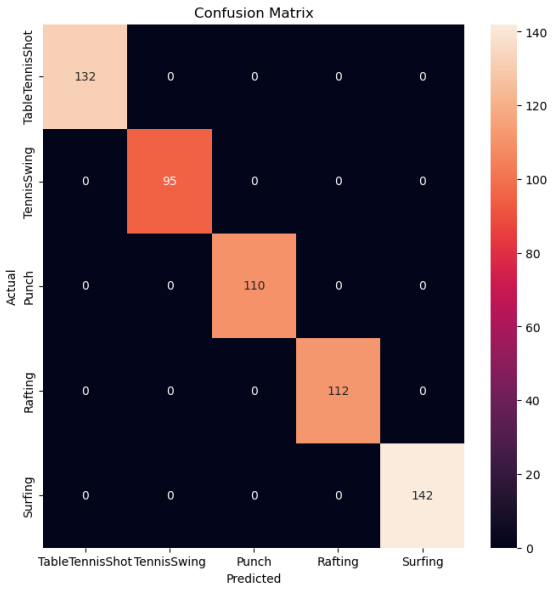
Burada ikinci seçilen resmin %1 oranında rafting kategorisine ait olduğunu %99 oranında sörf olarak doğru tahmin ettiği, diğer örnekler için de %100 oranında doğru tahmin ettiği görülmektedir. Masa tenisi kategorisine ait 4 farklı rastgele seçilen örneklere ait sonuçlar şe şekildedir:



Burada da tüm örnekleri mutlak kesinlikle doğru tahmin ettiği görülmektedir. Tenis raketi sallama kategorisine ait 4 farklı rasgele seçilen örneğe ait sonuçlar aşağıdaki gibidir:



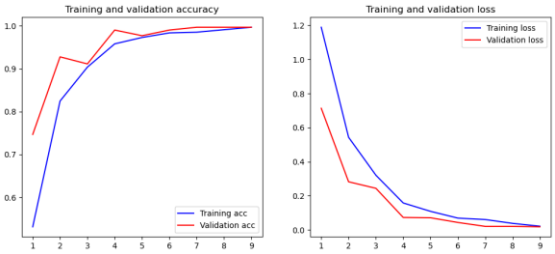
Burada da mutlak kesinlikle modelin doğru tahmin ettiği görülmektedir. 3D CNN modeline ait karmaşıklık matrisi aşağıdaki gibidir:



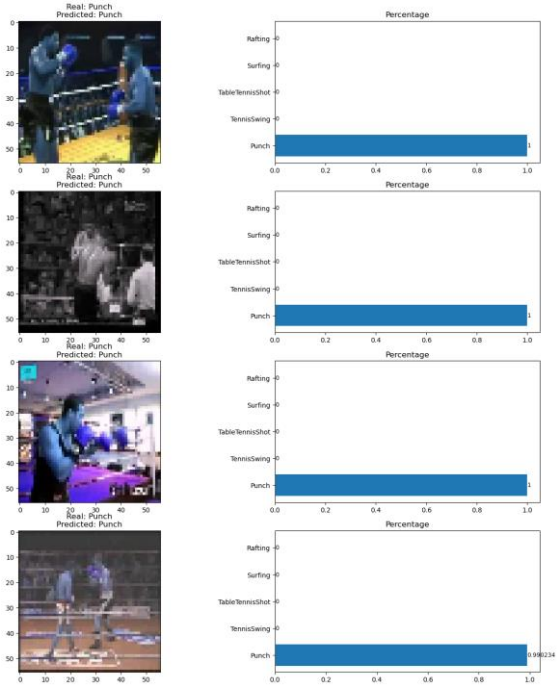
Karmaşıklık matrisine bakarak modelin mükemmel bir şekilde öğrendiği görülmektedir. Hiçbir test örneği için yanlış tahminde bulunmadığı görülmektedir.

CNN+LSTM Modeli

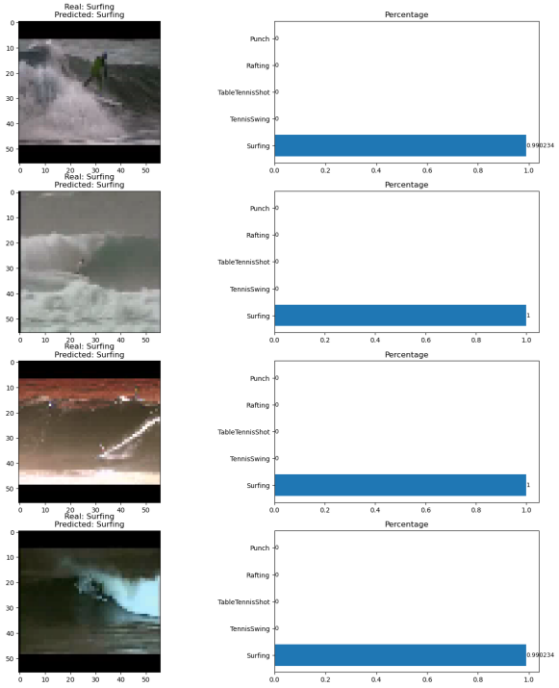
En iyi parametrelere sahip CNN+LSTM ağının eğitim grafiği aşağıdaki gibidir:



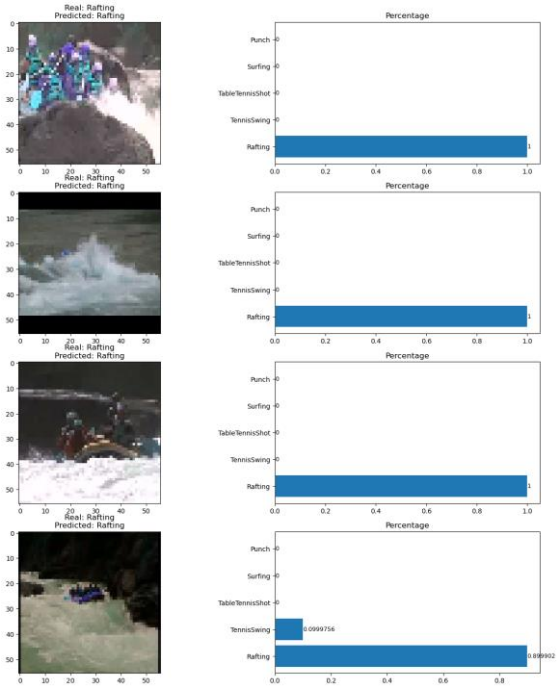
Burada eğitim ve validasyon başarılarının 1' e geldiği görülmektedir. Eğitim süresi 84,85 saniye sürmüştür. Test verisiyle test edildiğine başarı oranı %99.8 bulunmuştur. Punch kategorisine ait 4 farklı rastgele örnekle elde edilen sonuçlar şu şekildedir:



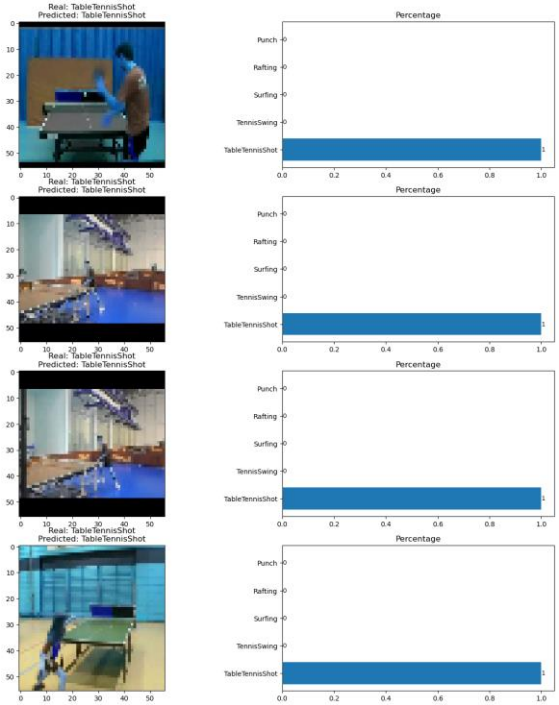
4 farklı örneğe baktığımızda da modelin kesin olarak doğru bildiği görülmektedir. Rafting kategorisine ait 4 farklı rastgele seçilen örneğe ait sonuçlar şu şekildedir:



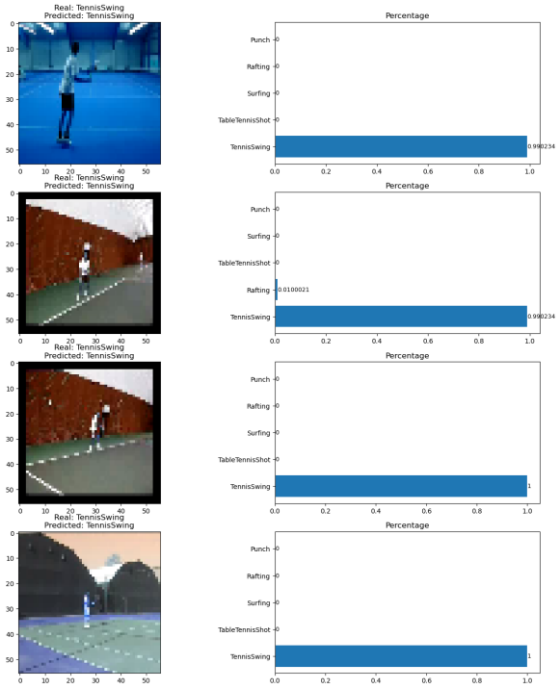
Burada 4 örneği de kesin olarak doğru tahmin ettiği görülmektedir. Masa tenisi şutu kategorisine ait 4 farklı rastgele seçilen örneklere ait sonuçlar şe şekildedir:



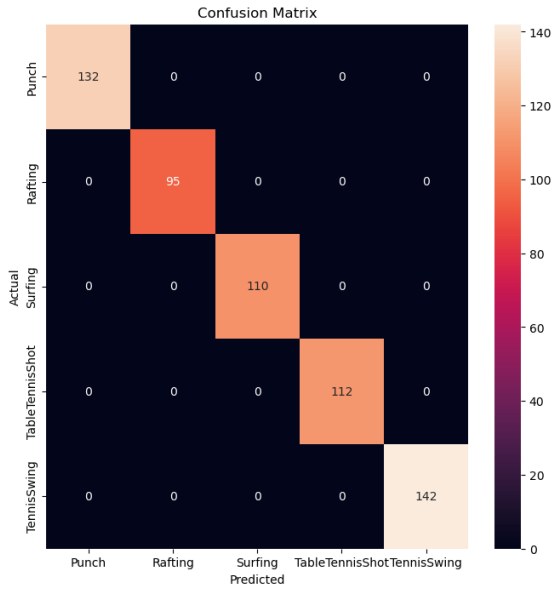
Burada son resim hariç diğer resimleri kesin olarak doğru tahmin ettiği son resimde ise tenis raketi sallama kategorisinin %10' luk bir dorğuluk payı olduğu görülmektedir. Sörf kategorisine ait 4 farklı rastgele seçilen resim aşağıdaki gibidir:



Burada da modelin kesin olarak doğru tahminde bulunduğu görülmektedir. Tenis raketi sallama kategorisine ait 4 farklı rasgele seçilen örneğe ait sonuçlar aşağıdaki gibidir:



Burada da mutlak kesinlikle modelin doğru tahmin ettiği görülmektedir. CNN+LSTM modeline ait karmaşıklık matrisi aşağıdaki gibidir:



Karmaşıklık matrisine bakarak modelin mükemmel bir şekilde öğrendiği görülmektedir. Hiçbir test örneği için yanlış tahminde bulunmadığı görülmektedir.

Sonuç

Hiper parametrelerin başarı üzerindeki etkisine baktığımızda aktivasyon fonksiyonu Sigmoid ve Tanh seçildiğinde en düşük başarıya sebep olduğu görülmektedir. Buna Vanishing gradient probleminin neden olduğu düşünülmektedir. Bu sorun, geriye doğru yayılım sırasında, ağın ilk katmanlarına doğru gradyanın giderek azalması veya sıfıra yaklaşması durumunda ortaya çıkar. Bu durum, bu katmanların güncellenmesi ve öğrenmesi zorlaştırır veya imkansız hale getirir.

Her iki modelinde çok iyi başarı oranları ortaya koyduğu görülmektedir. Aynı çerçeve sayısına sahip 3D CNN modellerinin CNN+LSTM modellerinden daha kısa sürede eğitimini tamamladığı görülmüştür. Bunun nedeninin LSTM doğası gereği daha fazla hesaplama karmaşıklığına sahip olması olduğu görülmüştür. Eğer iki modelden biri seçilecekse, eğitim ve test sürelerinin daha az olmasından dolayı 3D CNN daha makul bir tercih olacaktır.