Detekcija znakovnog jezika korišćenjem LSTM i MediaPipe

Uvod

Ovaj projekat ima za cilj razvoj AI sistema koji koristi LSTM (dugoročnu kratkoročnu memoriju) i MediaPipe tehnologiju kako bi prepoznao gestove znakovnog jezika. Znakovni jezik je ključna forma komunikacije za osobe sa oštećenim sluhom, ali razumevanje i tumačenje tih gestova može biti izazovno za druge ljude. Kroz primenu AI tehnika, želimo da prevaziđemo ovu prepreku i omogućimo lakšu i efikasniju komunikaciju sa osobama koje koriste znakovni jezik.

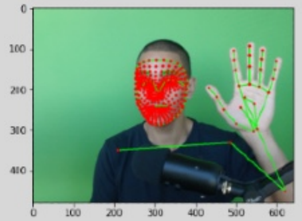
Prikupljanje podataka

Podatke o video snimcima znakovnog jezika smo prikupili iz WLASL skupa podataka. Metapodaci o video snimcima, uključujući oznake reči i informacije o instancama, su dobijeni iz WLASL fajlova.
Nedostajući video snimci su filtrirani koristeći unapred definisanu listu ID-eva nedostajućih video snimaka.

Izdvajanje podataka

Koristili smo MediaPipe-ov holistički model za izdvajanje ključnih tačaka iz svakog video kadra.
Izdvojene su ključne tačke za poziciju, lice, levu ruku i desnu ruku.
Kadrovi su sačuvani kao NumPy nizovi, pri čemu svaki niz predstavlja ključne tačke jednog kadra.

Ekstrakcija keypoint-a



Priprema podataka za obuku

Kreirane su fascikle za reči radi organizacije izdvojenih kadrova.
Kadrovi i oznake su učitani iz fascikli sa podacima za obuku.
Kadrovi su proširени (padovani) radi postizanja konzistentne dužine za ulaz u LSTM model.
Oznake su enkodirane u one-hot formatu radi kategorizacije.

Arhitektura LSTM modela

Konstruisan je sekvencijalni LSTM model. Model se sastoji od više LSTM slojeva sa sve većim brojem skrivenih jedinica.
Na svaki LSTM sloj primenjena je ReLU aktivaciona funkcija.
Dodati su gusti slojevi sa ReLU aktivacionom funkcijom radi transformacije karakteristika.
Konačni gusti sloj koristi softmax aktivacionu funkciju za višeklasnu klasifikaciju.

Obuka modela

LSTM model je kompajliran sa Adam optimizatorom i gubitkom kategoričke unakrsne entropije.
Obuka je izvršena tokom 20 epoha sa podacima za obuku.
TensorBoard je korišćen za vizualizaciju napretka obuke i metrika performansi.

Problem različite dužine frame-ova

Jedan od izazova u obradi video zapisa za prepoznavanje znakovnog jezika je različita dužina frame-ova u različitim video snimcima. LSTM model, međutim, očekuje podatke fiksne dužine kao ulaz. Da bismo rešili ovaj problem, primenjujemo tehniku proširenja (padovanja) podataka. Proširenje podataka se vrši tako što se kraći frame-ovi dopunjavaju nulama ili se seku duži frame-ovi kako bi svi bili iste dužine. Ova tehnika omogućava LSTM modelu da efikasno obradi video zapise različitih dužina i izvrši prepoznavanje znakovnog jezika.

Evaluacija modela

Testni kadrovi su prosleđeni obučenom LSTM modelu radi predviđanja.
Predviđene oznake su upoređene sa stvarnim oznakama radi izračunavanja tačnosti.
Izračunate su matrica konfuzije sa višestrukim oznakama i rezultat tačnosti radi procene performansi modela.

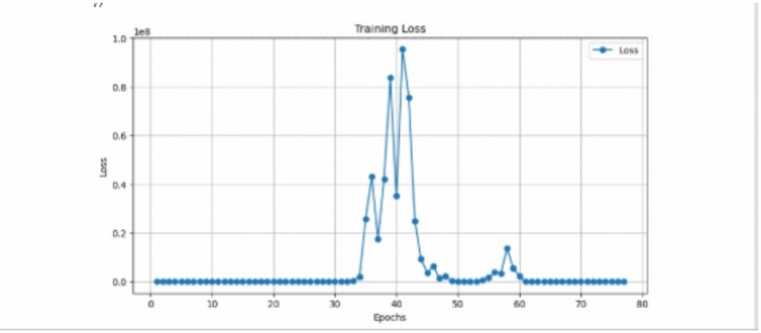
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, None, 64)	442112
lstm_1 (LSTM)	(None, None, 128)	98816
lstm_2 (LSTM)	(None, 64)	49408
dense (Dense)	(None, 64)	4160
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 1069)	35277

Total params: 631853 (2.41 MB)
Trainable params: 631853 (2.41 MB)
Non-trainable params: 0 (0.00 Byte)

Rezultati testiranja

Ispod su prikazani rezultati na 3 reči i 77 epoha.
Na prvom grafu je prikazarn loss po epohama, a na drugom categorical accuracy.
Accuracy score je 1.0.

```
In [23]: words_predicted = model.predict(frames_array_test)
         multilabel_confusion_matrix(np.argmax(labels_array_test, axis=1), np.argmax(words_predicted , axis=1))
         accuracy_score(np.argmax(labels_array_test, axis=1), np.argmax(words_predicted , axis=1))
         1/1 [==============================] - 0s 297ms/step
Out[23]: 1.0
```



Fakuktet tehiničkih nauka, Novi Sad
4. jul 2023.
Danilo Babić i Jelena Miković