VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Strojové učenie a rozpoznávanie

Projekt

4. května 2025

Bc. Lucia Šírová, Bc. Daniel Blaško

Klasifikovanie hlasu

Ide o systém, ktorý umožňuje identifikovať, ktorá osoba z množiny známych hovoriacich vystupuje v danej zvukovej nahrávke. Implementovaný systém je založený na Gaussian mixture modeli (GMM), pričom pre každého hovoriaceho je trénovaný samostatný model s 32 Gaussovskými komponentami. Model bol trénovaný prostredníctvom EM algoritmu na príznakoch Mel-Frequency Cepstral Coefficients (MFCC) extrahovaných z jednotlivých zvukových nahrávok.

Extrakcia príznakov

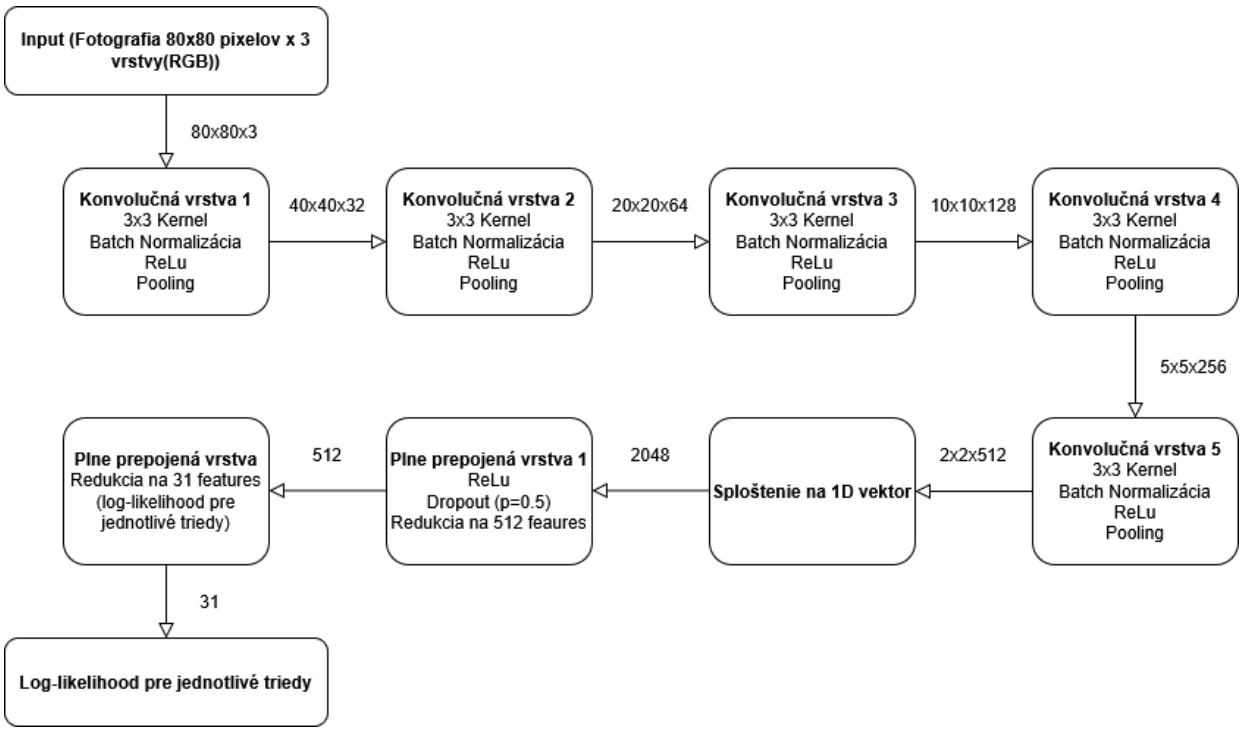
Príznaky boli z rečových nahrávok extrahované pomocou funkcie `wav16khz2mfcc`, ktorá vychádza z knižnice `ikrlib`. Na zvýšenie presnosti klasifikátora bola funkcia rozšírená o filtrovanie segmentov s nízkou energiou, čím sa eliminujú tiché časti nahrávky. Taktiež bola realizovaná Cepstrálna stredná normalizácia (CMN) na odstránenie skreslenia nahrávok spôsobených rozdielnymi podmienkami pri nahrávaní.

Vyhodnocovanie systému

Počas ladenia parametrov GMM modelu a úprav predspracovania hlasových nahrávok bol klasifikátor trénovaný na dátach zo zložky `train` a testovaný na dátach zo zložky `dev`. Úspešnosť systému bola vyhodnotená ako podiel správne klasifikovaných nahrávok k celkovému počtu testovaných nahrávok.

Klasifikovanie obrázkov

Tento systém, rovnako ako v predošlej sekcii, identifikuje, o ktorú z 31 osôb sa jedná, tentokrát však z fotografie tváre. Model je založený na konvolučnej neurónovej sieti, jednotnej pre všetky osoby. Model má 2635551 parametrov. Neurónová sieť má nasledujúcu štruktúru:



Obrázek 1: Diagram Konvolučnej neurónovej siete pre klasifikáciu tvárí

Trénovanie modelu

Pre zisťovanie najlepších parametrov modelu som využil jackknifing, kedy som použil jeden obrázok z každej triedy z množiny validačných dát, keďže išlo o ďalšie sedenie (Session 4), čo znamenalo väčšiu rozmanitosť trénovacích dát. Vo validačnej sade boli nahradené jednou vzorkou zo Session 2 z trénovacej sady. Tento prístup naozaj poznateľne zvýšil kvalitu modelu oproti štandardnému rozdeleniu dát.

Množina trénovacích dát je však veľmi malá, preto bolo potrebné vykonať augumentáciu dát pre lepšiu schopnosť modelu generalizovať. Toto je možné vykonať dvoma spôsobmi. Prvým je offline augumentácia, kedy by boli trénovacie dáta vopred augumentované a uložené, a až neskôr spracované. Nárast veľkosti trénovacích dát by z dôvodu veľmi malého datasetu nepredstavoval problém. Druhým spôsobom je online augumentácia, kedy sa pri každej epoche trénovania vstupné dáta augumentujú náhodnou podmnožinou množiny transformácií. Rozhodol som sa implementovať online augumentáciu, keďže poskytuje väčšiu variabilitu rôznych transformácií (vykonať a uložiť všetky kombinácie podmožín transformácií by bolo príliš náročné). Zoznam transformácií je nasledovný:

- Geometrické transformácie
  - Horizontálne otočenie (p=0.5)
  - Rotácia (-10°,10°, p=0.5)
  - Posun (10%, p=0.5)
  - Mierka (90%, 110%, p=0.5)
  - Skosenie (-1°, 1°, p=0.5)
  - Orezanie (mierka (0.9, 1.0), pomer (0.9, 1.1))
  - Transformácia perspektívy (mierka=0.05, p=0.5)
- Farebné transformácie
  - Zmena farieb (jas=0.2, kontrast=0.2, saturácia=0.1, odtiene=0.02)
- Šum, ostrosť
  - Gaussovské rozostrenie (veľkosť kernelu=3, σ=(0.1, 0.5))
  - Gaussovský šum (stredná hodnota=0, odchýlka=0.01, p=0.5)
  - Úprava ostrosti (faktor=1.5, p=0.5)
  - Autokontrast (p=0.3)
- Normalizácia RGB kanálu
  - Normalizácia (stred=[0.5072, 0.3863, 0.3990], odchýlka=[0.1851, 0.1923, 0.1830])

Funkcia pre pridanie gaussovského zašumenia musela byť doimplementovaná. Hodnoty pre normalizáciu RGB kanálov boli vypočítané z zo všetkých dostupných dát. Tieto hodnoty sú relatívne blízke bežne používaným hodnotám získaným z ImageNet databázy (stred=(0.485, 0.456, 0.406), odchýlka=(0.229, 0.224, 0.225)).

Pri validácii modelu bol na časť validačných dát aplikovaný gaussovský šum s cieľom simulovať schopnosť modelu správne klasifikovať aj pri zašumených obrázkoch. Úspešnosť dosahuje veľmi podobných hodnôt v porovnaní s neupravenými validačnými dátami.

Pre trénovanie modelu boli vybrané nasledovné parametre:

- Batch size: *16*
- Criterion: *CrossEntropyLoss*
- Optimizér: *Adaptive Moment Estimation*
- Learning Rate: *0.01*
- Počet epoch: *30* (pri vyššom počte začal model vykazovať známky pretrénovania)

Spojenie klasifikátorov

Vrámci snahy o vylepšenie presnosti klasifikácie bola vytvorená fúzia klasifikátora hlasu a obrazu. Každý z vyššie uvedených klasifikátorov je skriptom `combined_classifier.py` spustený samostatne. Výsledky z

oboch klasifikátorov sú spracované a hodnoty logaritmickej vierohodnosti sú normalizované spojené váženým priemerom:

$$score_{fused} = \alpha \cdot score_{voice} + (1 - \alpha) \cdot score_{image}$$

Kde α je váha hlasového klasifikátora. Finálna klasifikácia bola určená ako trieda s najvyšším skóre po fúzii.

Použitie

Zložka **/src** je rozdelená do podzložiek: **/eval_scripts** a **/train_scripts**. Zložka **/eval_scripts** obsahuje verziu skriptov, ktoré načítavajú testovacie dáta vo formáte zodpovedajúcom ostrému testovaniu našich modelov pre hodnotenie (súbory nie sú rozdelené podľa triedy). Zložka **/train_scripts** obsahuje verziu skriptov, ktorá očakáva dáta vo formáte, ktorý vychádza zo zadaných zložiek **/train** a **/dev** (súbory sú rozdelené podľa triedy). V adresároch sa kvôli veľkostnému limitu pri odovzdaní nenachádzajú súbory na trénovanie a testovanie. Keďže boli pri trénovaní klasifikátora na rozpoznávanie obrázkov zložky testovacích a trénovacích dát modifikované, prikladáme odkaz na cloudové úložisko, kde sa tieto zložky nachádzajú, spolu s natrénovaným modelom: Link

Pred používaním systémov na rozpoznávanie osôb je nutné inštalovať potrebné závislosti použitím príkazu:
`pip3 install -r src/requirements.txt`

Klasifikátor hlasu

Systém na klasifikovanie osoby na základe hlasu je možné spustiť nasledovne:
`python3 voice_classfier.py --train_dir <train_dir> --test_dir <test_dir>`
Súbor s výsledkami klasifikácie testovacích dát je uložený pod názvom **voice_results.txt**.

Klasifikátor fotografií

Trénovanie modulu je spustené príkazom `python3 image_model_train.py --train_dir [directory] --epochs [num_epochs]`, kde `--train_dir` určuje zložku s trénovacími dátami (default je **train/**) a `--epochs` určuje počet epoch trénovania modelu (default je 30).

Validáciu natrénovaného modelu je možné spustiť príkazom `python3 image_validation.py --val_dir [directory]`, kde `--val_dir` určuje zložku s validačnými dátami (default je **dev/**).

Spustenie modelu nad testovacími dátami je vykonané príkazom `python3 image_classifier.py --test_dir [directory]`. Default zložka je **test/** a výsledky sú uložené do **image_results.txt**.

Spojený klasifikátor

Systém na klasifikovanie osoby na základe hlasu a obrazu je možné spustiť nasledovne:
`python3 combined_classfier_eval.py --train_dir_image <train_dir> --train_dir_voice <train_dir> --test_dir <test_dir> --train_img_model`. Posledný prepínač slúži na nové natrénovanie modelu, ak ešte model natrénovaný nebol (kvôli limitu veľkosti nebolo možné model priložiť do archívu s riešením, keďže jeho veľkosť je cca. 10MB).
Súbor s výsledkami klasifikácie testovacích dát je uložený v koreňovom adresári pod názvom **combined_results.txt**.

Výsledky pre jednotlivé modely obsahujú log-likelihood hodnoty jednotlivých tried, avšak kvôli použitiu rozdielnych modelov bolo nutné jednotlivé log likelihood hodnoty normalizovať, a teda v kombinovanom klasifikátore nie sú dostupné hodnoty log-likelihood hodnoty. Z toho dôvodu sú v súbore s výsledkami hodnoty NaN.