Vysoké učení technické v Brně
Fakulta informačních technologií

Projekt do předmětu SUR

Klasifikace obrazu a řeči

Venkrbec Tomáš (xvenkr01)
Vosol David (xvosol00)
Willaschek Tomáš (xwilla00)

14. května 2021

1 Úvod

V tomto projektu je naším cílem vytvořit dva klasifikátory, které se zaměřují na úlohu rozpoznání osoby na fotografii a rozpoznání mluvčího na zvukové nahrávce. Pro obě tyto úlohy jsme zvolili řešení pomocí neuronových sítí.

2 Datová sada

Zvukové nahrávky a fotografie z několika nahrávacích sezení celkem 31 osob, převážně zaměstnanců VUT FIT.

Obraz: PNG formát, Rozlišení 80x80 pixelů
Zvuk: WAV formát, Kanál Mono, Vzorkovací frekvence 16 000Hz

3 Zpracování zvuku

Pro zpracování zvukového záznamu byla využita konvoluční neuronová síť. Jejímž vstupem je zvukový záznam o délce 1s. Tedy při vzorkovací frekvenci 16KHz je to vektor o rozměru [16000,1].

3.1 Příprava dat

Záznam vznikl rozdělením původní nahrávky z datasetu tak, aby každý jeden vzorek pro neuronovou síť byl stejné délky. Případné kratší části jsou nulami zarovnány na požadovanou délku 1s. Ještě před tímto rozdělením bylo potřeba data očistit. Byla vytvořena obálka, jejíž úlohou bylo ořezání frekvencí a částí nahrávky, kde se nenachází samotný hlasový záznam dané osoby, tedy typicky začátek a konec záznamu, případně delší promlky uprostřed nahrávky, které bývají většinou vyplněny tichem. Empiricky bylo zjištěno, že síť nejlépe pracuje se vzorky, které byly ořezány pod hodnotou 600.

Byly zde také pokusy o větší využití knihovny *Librosa*[2], tedy zejména prvotní očištění nahrávek, kdy jsme opět chtěli ořezat úvodní a koncové ticho, vyextrahovat hlas osob od pozadí pomocí `nn_filter`, využít *STFT*, či normalizovat hodnoty `wav` souborů do $\langle -1, 1 \rangle$. Ačkoliv všechny tyto úpravy byly poměrně výpočetně náročné, naše navržená architektura neuronové sítě s nimi nějakým způsobem nedokázala správně naložit, a tak dosahovala velice špatných výsledků s přesností pohybující se kolem 30% na validační sadě. Od těchto přístupů bylo tedy upuštěno a dále se pokračovalo pouze s naší jednoduchou ořezávací obálkou zmíněnou výše.

3.2 CNN

Jako první vrstva sítě je `get_melspectrogram_layer`, která je implementována v knihovně *Kapre* [1]. Ta pomáhá zjednodušit práci se zvukovými záznamy při strojovém učení tak, že nyní můžeme mít samotné získávání *MFCC* parametrů rovnou součástí neuronové sítě. Na vstup tedy můžeme dát náš upravený záznam a pomocí této vrstvy rovnou vytvořit Mel spektrogramy. (Není potřeba tedy nejprve počítat *Fourierovu transformaci*, počítat *Mel filter bank* koeficienty atd. [3].) Na tuto vrstvu poté navazují klasické *2D konvoluční a maxpooling* vrstvy, které v zásadě nejsou nijak extra zajímavé a nalezneme je v každé síti zpracovávající obrázky. Jako poslední vrstva v síti je vrstva plně propojená s 31 neurony na výstupu a funkcí *Softmax*, kde tedy dostáváme normalizované pravděpodobnosti jednotlivých tříd. Tyto pravděpodobnosti byly převedeny do logaritmické domény pro pozdější manipulaci. Neuronová síť je v tomto případě implementována ve frameworku `Tensorflow` s využitím knihovny `Keras`.

Jelikož trénink probíhal na vzorcích o délce 1s, ale nahrávky jednotlivých osob jsou podstatně delší a je jich také větší množství, bereme jako celkový výsledek klasifikace tu třídu, jejíž celková pravděpodobnost ze všech vzorků je nejvyšší. Předpokládáme tedy jednotlivé 1s vzorky jako podmíněně nezávislé jevy, a proto si můžeme dovolit tyto pravděpodobnosti jednoduše pronásobit mezi sebou a získat tak výslednou log pravděpodobnost jednotlivých tříd přes všechny 1s vzorky dané osoby.

3.3 Vyhodnocení

Po poměrně extenzivním hledání vhodných hodnot hyperparametrů, bylo nakonec dosaženo úspěšnosti 76%, což osobně nehodnotím moc kladně. Myslím si, že to může být způsobeno zejména malým množstvím trénovacích vzorků a teoreticky jejich nedostatečným očistěním, ačkoliv při vyšším očistění dosahovala síť daleko horších výsledků.

Podle mého názoru se na tuto úlohu spíše hodí jiné mechanismy typu SVM, případně GMM, které nepotřebují takové množství vstupních dat, jako neuronové sítě. Dále by bylo možné uměle vytvořit větší množství vzorků, buď zkrácením jejich délky, nebo jejich augmentací. V neposlední řadě by stálo za hlubší prozkoumání a využití i nějakého bloku typu *LSTM*, ačkoliv při prvotních testech se mi tato komponenta příliš neosvědčila.

4 Zpracování obrazu

Pro klasifikaci osob na základě fotografií byla využita konvoluční neuronová síť. Vstupní obrázky z datové sady nebylo zapotřebí nijak upravovat, pouze byly při načítání hodnoty jednotlivých pixelů normalizovány do rozsahu $< -1; 1 >$, který je pro neuronovou síť vhodný. Malá velikost datové sady neumožňovala použití příliš velké a složité sítě, které by se snadněji přeučily na trénovacích datech, proto byla použita síť architektury `VGG` s 3 konvolučními bloky, zakončená plně propojenou vrstvou s aktivační funkcí *softmax*, vracející pravděpodobnosti všech 31 tříd. Ve všech konvolučních vrstvách byla použita *dávková normalizace*, společně s aktivační funkcí *LeakyReLU*, které se po různých experimentech s nastavením sítě projevily jako nejvhodnější. Nedílnou součástí k zlepšení regularizace na malé sadě je také vrstva *Dropout*, která je použita za každým konvolučním blokem. Trénování probíhá s pomocí optimalizátoru *Adam* a při trénování je minimalizována chybová funkce *kosinová podobnost*.

4.1 Augmentace datové sady

Jak již bylo zmíněno v předchozí sekci, limitujícím faktorem použití konvolučních sítí v této úloze byla právě velmi malá velikost datové sady. Tento problém byl u zpracování obrázků zmírněn použitím náhodných transformací na jednotlivé načítané vzorky. Všechny vzorky v každé trénovací dávce jsou tedy náhodně posunovány vertikálně i horizontálně, jsou rotovány, přibližovány či oddalovány a také je náhodně upravován jejich jas.

4.2 Vyhodnocení

Stejně jako konvoluční síť pro zpracování zvuku, i tato síť vyžadovala k dosažení uspokojující hodnoty validační přesnosti spoustu mnoho práce s nastavením hyperparametrů a úpravami struktury sítě. V případě klasifikace osob na základě fotografií se podařilo dosáhnout přesnosti 75.8%. Dosažení vyšší přesnosti je komplikováno tím, jak snadno dochází na takto malé trénovací datové sadě k přeučení, i s použitím různých regularizačních technik.

5 Spuštění

Knihovny potřebné ke spuštění programu jsou vypsány v souboru `requirements.txt` a jsou stažitelné obvyklým způsobem přes instalátor balíčků `pip`. Spouštěč `run.py` má vícero parametrů. Pro jejich výpis s vysvětleném stačí použít přepínač `-h`.

Příklad spuštění:

```
# spuštění trénování obou sítí
python3 run.py --run_all --test_path PATH --val_path PATH \
    --train_path PATH  --train

# spuštění evaluace obou neuronových sítí
python3 run.py --run_all --test_path PATH --val_path PATH \
    --train_path PATH
```

Při spuštění s `-train` skript vytvoří `.h5` soubory v adresáři `snapshots`. Každý pro daný klasifikátor. Při spuštění evaluace vytvoří skript dva soubory `image_classifier.txt` a `voice_classifier.txt` obsahující výsledky. Oba v rootu adresáře (stejná úroveň se `SRC`).

6 Závěr

Tento projekt byl zajímavý, protože někteří z členů týmu se doposud osobně nesetkali se zpracováním zvuku či obrazu. Díky tomuto projektu jsme aplikovali či prozkoumali nové praktiky, což bude mít jistě přínos pro budoucí implementace. Oba vytvořené modely se nám podařilo natrénovat s poměrně dobrou přesností vzhledem k dostupnému počtu dat.

Reference

[1] *Kapre - Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras*. Dostupné z: https://kapre.readthedocs.io/en/latest/.

[2] *Librosa: Audio and music signal analysis in python*. Dostupné z: https://librosa.org/doc/latest/index.html.

[3] BURGET, L. *Slajdy do předmětu SUR - Extrakce příznaků*. 2021. Dostupné z: https://www.fit.vutbr.cz/study/courses/SUR/public/prednasky/03_extrakce_priznaku.