DEEP LEARNING BASED SOUND EVENT RECOGNITION

Jakub Bajzík
Master Degree Programme (3), FEEC BUT
E-mail: xbajzi00@stud.feec.vutbr.cz

Supervised by: Jiří Přinosil
E-mail: prinosil@feec.vutbr.cz

Abstract: The main paper deals with the analysis of the methods of processing and recognition of events in the audio signal and the implementation of the selected method in real use. Recognized events are gunshots placed in a background sound such as traffic noise, human voice, animal sounds and other forms of environmental sounds. For events classification and class recognition, the freely available machine learning framework TensorFlow is used.

Keywords: Sound recognition, machine learning, neural network, signal processing

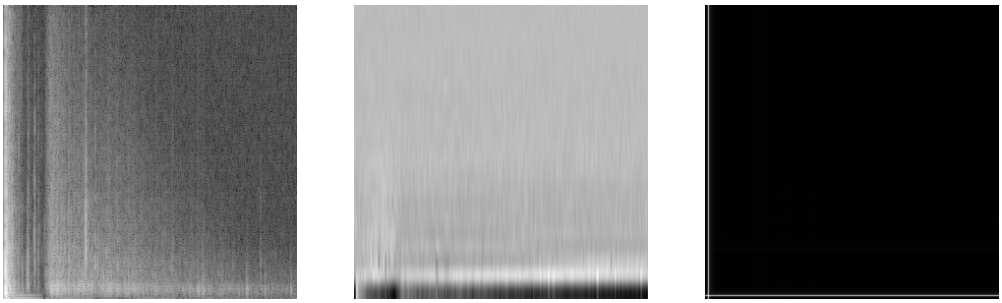
1 ÚVOD

Rozpoznávanie objektov pomocou strojového učenia je najčastejšie spájané s obrazovým signálom. V prípade spracovania zvukových signálov mimo hudobných sú dnes dobre známe najmä metódy spracovania ľudského hlasu. Často sa však skloňuje použitie známych postupov pre rozpoznanie zvukových udalostí okolitého prostredia, ktoré môžu byť výbuch, výstrel zo zbrane, siréna, poplašné zariadenie auta, detský plač, rozbitie okna a iné udalosti spájané s potenciálnym nebezpečenstvom. Využitie takto naučených algoritmov je najmä zvýšenie bezpečnosti majetku a osôb. Implementácia je možná napríklad v domových alebo priemyselných systémoch ochrany. Obsah práce je zameraný na možnosť využitia vizualizácie zvuku ako príznaku pre učenie konvolučnej neurónovej siete na rozpoznanie výstrelu zo strelnej zbrane od náhodného pozadia. Okrem často používaného spektrogramu sú hľadané nové vizualizácie, prípadne kombinácia viacerých s najvyššou úspešnosťou predikcie v reálnom prostredí.

2 VIZUALIZÁCIA ZVUKU

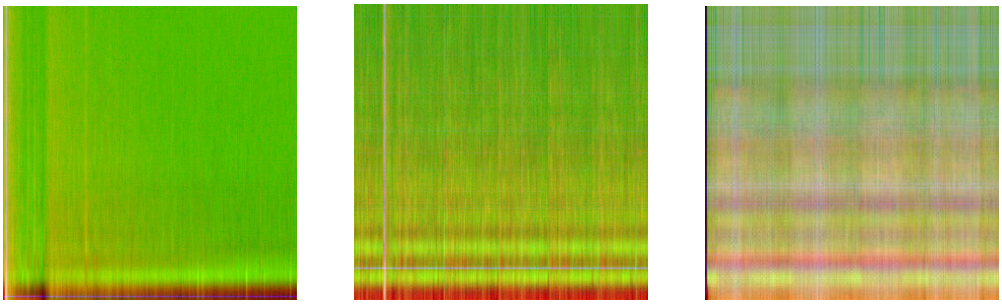
Najpoužívanejšou vizualizáciou zvukového signálu je spektrogram, ktorý zobrazuje vývoj frekvenčného spektra v čase. Pre prevod signálu z časovej do frekvenčnej oblasti je použitá krátkodobá fourierová transformácia FFT [1]. Ďalšou použitou reprezentáciou sú melovské kepstrálne koeficienty MFCC ktoré vychádzajú z nelineárnych a maskovacích vlastností ľudského sluchu. Násobením spektra signálu nelineárne rozloženou bankou filtrov, logaritmovaním a následnou spätnou diskrétnou kosínovou transformáciou získame koeficienty MFCC [2]. Posledným použitým príznakom je miera vlastnej podobnosti založená na vzdialenostiach. Táto technika sa používa na analýzu globálnej štruktúry hudobných diel. Pre zobrazenie použijeme maticu vlastnej podobnosti. Vertikálna a horizontálna os zobrazenia predstavuje časovú postupnosť. Najväčšia podobnosť je na hlavnej diagonále, podľa ktorej je matica súmerná [3]. Takto vzniknuté dvojrozmerné obrazce zobrazené na obrázku 1 boli použité ako jednotlivé RGB kanály výsledného obrazu, ktorý prechádza neurónovou sieťou.

Tento koncept bol použitý v práci [4] pre klasifikáciu enviromentálnych zvukov. Výsledkom bolo zistenie, že úspešnosť klasifikácie nezvýši použitie matice podobnosti a MFCC koeficientov v porovnaní so samotným spektrogramom. V prípade rozpoznávania zvuku výstrelu od náhodného pozadia existuje predpoklad, že tieto dva prídavné obrazce odhalia zvuky s odlišným charakterom od výstrelov.



(a) Spektrogram (b) MFCC koeficienty (c) Matica podobnosti

Obrázek 1: Jednotlivé farebné kanály RGB vizualizácie výstrelu.



(a) Výstrel zo zbrane (b) Výstrel a pozadie (c) Ruch ulice

Obrázek 2: Výsledná vizualizácia zvukov ako RGB obraz.

Po vykreslení matice podobnosti výstrelu vidíme len dve úzke čiary, no v prípade zvuku s pravidelnou periodicitou sa v obraze začne objavovať pravidelná mriežka. MFCC koeficienty môžu odhaliť zvuky rečového charakteru.

3 DATABÁZA NAHRÁVOK

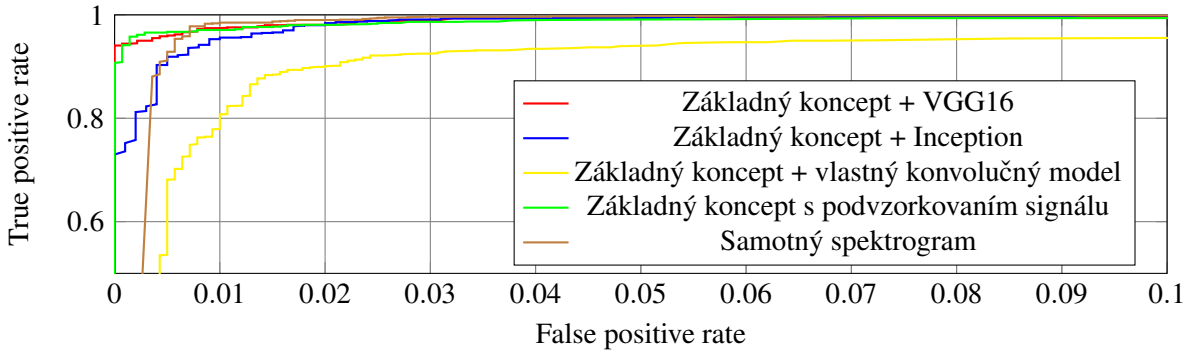
Zvukové nahrávky výstrelov a náhodných pozadí boli získané z voľne dostupných zvukových databáz. Databáza výstrelov má názov *The Free Firearm Sound Library – Expanded Edition* a obsahuje viac ako 1000 sekundových nahrávok. Databáza nahrávok pozadí *UrbanSound8K* obsahuje viac ako 8000 zvukov z rôznych zdrojov najmä hluk ulice, detský krik, štekanie psov, sirény a iné hlukové pozadie. Pôvodne obsahovala tiež výstrely zo zbraní, tie však boli odstránené. Nahrávky sú vo formáte WAV so vzorkovacou frekvenciou 44,1 kHz. Trénovaciu množinu tvorí 60%, validačnú 20% a testovaciu zostávajúcich 20% nahrávok.

4 NEURONOVÁ SIEŤ

Umelú neurónovú sieť tvorí vstupná vrstva, ktorá sprostredkúva spojenie so vstupnými dátami, skrytá vrstva a výstupná vrstva [1]. Na prevod dvojrozmerných dát slúži vstupná konvolučná vrstva, ktorej výstupom je vektor hodnôt a je priamo prepojená so skrytou vrstvou. Preto v tomto prípade hovoríme o konvolučnej neurónovej sieti. Výstupnú vrstvu tvoria v prípade binárnej klasifikácie dva neuróny, ktorých výstupy odpovedajú pravdepodobnosti zaradenia vstupných dát do príslušnej triedy, v tomto prípade či sa jedná o výstrel alebo pozadie. V práci sú porovnané modely VGG16, Inception a vlastný konvolučný model bez natrénovaných váh. V ostatných prípadoch je použitý model VGG16 kvôli najvyššej presnosti. Samotná sieť je zostavená pomocou frameworku TensorFlow s nadstavbou Keras.

5 EXPERIMENT A VÝSLEDKY

Experiment pozostáva z viacerých testov, ktoré porovnávajú rôzne konvolučné modely a postupy vizualizácie. V jednom z testov bol signál pred spracovaním podvzorkovaný na 8 kHz. Pre výpočet spektrogramu a MFCC bol signál najskôr rozdelený na segmenty s dĺžkou 256 vzorkov s polovičným prekrytím a váhovaný hammingovým oknom. Následne bolo vypočítaných 4096 koeficientov FFT a 20 koeficientov MFCC. Pred výpočtom matice podobnosti bol signál podvzorkovaný na 10 kHz. Pre vyhodnotenie úspešnosti predikcie použijeme ROC [5]. V tomto prípade pozitívny výsledok znamená predikovaný výstrel, negatívny náhodné pozadie.



Obrázek 3: Porovnanie úspešnosti predikcie pomocou ROC kriviek.

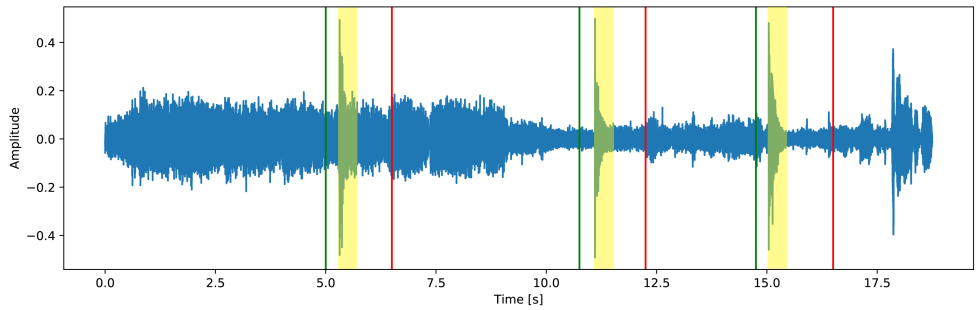
Na priebehu vidno, že červená a zelená krivka sa takmer zhodujú, čo znamená, že zredukovaním dátového toku podvzorkovaním sa nezníži úspešnosť rozpoznania. Rozdiel medzi hnedou krivkou predstavujúcou model natrénovaný na samotný spektrogram a červenou krivkou konceptu s troma vrstvami je viditeľný najmä v porovnaní miesta poklesu. Hnedá krivka začína klesať skôr ako červená, čo znamená menej falošných výstrelov pri rovnakej presnosti predikcie skutočných výstrelov. Najhoršie výsledky dosahuje vlastný konvolučný model. Následná tabuľka 1 zobrazuje opäť porovnanie úspešnosti modelu na základe plochy pod ROC krivkou, celkovej presnosti modelu a presnosti rozpoznania výstrelov (zhodnosť). Výsledky odpovedajú rovnako nastavenej chybovosti, aby s takto nastaveným prahom bolo najviac 0,5% pozadí nesprávne vyhodnotených ako výstrel. Pokiaľ by sme požadovali ešte nižšiu chybovosť, presnosť na základe samotného spektrogramu výrazne klesne.

Tabulka 1: Porovnanie úspešnosti rozpoznania v jednotlivých testoch.

Testovaný model	Plocha pod ROC	Presnosť	Zhodnosť
VGG16	0,9991	0,9771	0,9955
Inception	0,9980	0,9570	0,9946
Vlastný konvolučný model	0,9754	0,8318	0,9926
Podvzorkovaný signál	0,9988	0,9811	0,9949
Samotný spektrogram	0,9976	0,9525	0,9953

Výsledná aplikácia spracováva zvukový signál v sekundových intervaloch s posunom štvrtiny sekundy. Použitý je konvolučný model VGG16. Výstupom aplikácie sú označené úseky v ktorých bol nájdený výstrel, pričom zelená čiara označuje začiatok a červené koniec úseku. Na obrázku 4 je zobrazený časový priebeh testovanej nahrávky, ktorá obsahuje krik ľudí, zrážky áut a iné ruchy ku ktorým boli pridané tri výstreli zo strelných zbraní. Výstreli boli nahrané na strelnici v rámci práce za účelom priblíženia testovania k reálnym podmienkam. Skutočné pozície výstrelov sú označené žltou farbou. Na základe spektrogramu, MFCC a podobnosti aplikácia správne označila tri úseky a nepomýlila

ju zrážka dvoch áut ani impulzívny zvuk otvorenia dverí na konci nahrávky. Rovnaký výsledok bol dosiahnutý pri použití samotného spektrogramu.



Obrázek 4: Označené úseky s výstrelmi v testovacej nahrávke.

6 ZÁVER

V práci boli preskúmané možné vizualizácie zvukového signálu a ich vplyv na úspešnosť predikcie pomocou konvolučných neuronových sietí. Výsledky naznačujú, že spojenie spektrogramu, MFCC koeficientov a matice podobnosti môže viesť k menšiemu množstvu falošných predikcií výstrelu v porovnaní so samotným spektrogramom, čo sa však na testovanej nahrávke neprejavilo. Výhodou spektrogramu je nižší čas spracovania signálu, ktorý hrá úlohu najmä v prípade predikcie v reálnom čase na zariadeniach s menším výpočtovým výkonom. Experiment ukázal, že zvýšenie rýchlosti spracovania pri zachovanej presnosti je možné dosiahnuť zmenšením dátového toku podvzorkovaním signálu.

Výsledky práce potvrdzujú, že použitie natrénovaných konvolučných modelov VGG16 a Inception zvyšuje presnosť predikcie výstrelov napriek tomu, že tieto modely sú primárne určené na rozpoznávanie obrazu. Skutočná presnosť závisí najmä od zvoleného prahu rozhodovania, ktorý je vhodné nastaviť s ohľadom na rušnosť pozadia zvukového signálu. Pokiaľ je ruch pozadia výrazný a príliš premenlivý, môže byť vhodné na základe impulzného charakteru zvuku výstrelu doplniť rozpoznávací algoritmus o detektor aktivity vo vyšších frekvenčných pásmach a tým potenciálne znížiť počet falošných výstrelov. Práve návrh detektoru môže byť motiváciou pre ďalšie pokračovanie práce.

REFERENCE

[1] MASTERS, Timothy. *Signal and image processing with neural networks*. John Wiley & Sons, Inc, 1994. 399 s. ISBN 0-471-04963-8.

[2] SMÉKAL, Zdenek. *Číslicové zpracování řeči*. Skriptum Ústav telekomunikací VUT v Brne, posledná aktualizácia 2010. 134 s.

[3] FOOTE, Jonathan T., COOPER, Matthew L. *Media Segmentation using Self-Similarity Decomposition*. Publikové v SPIE Storage and Retrieval for Media Databases 2003, Vol. 5021, s. 167-175.

[4] BODDAPATI, Venkatesh, PETEF, Andrej, RASMUSSON, Jim, LUNDBERG, Lars. *Classifying environmental sounds using image recognition networks*. Publikované v Procedia Computer Science, 2017. s. 2048–2056.

[5] FAWCETT, Tom. *An introduction to ROC analysis*. Publikované v Pattern Recognition Letters 27 2005. s. 861-874.