Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Ilkovičova 2, 842 16 Bratislava 4

Neurónové siete

maCaPeLa

Semestrálny projekt

Členovia tímu: Bc. Adam Puškáš

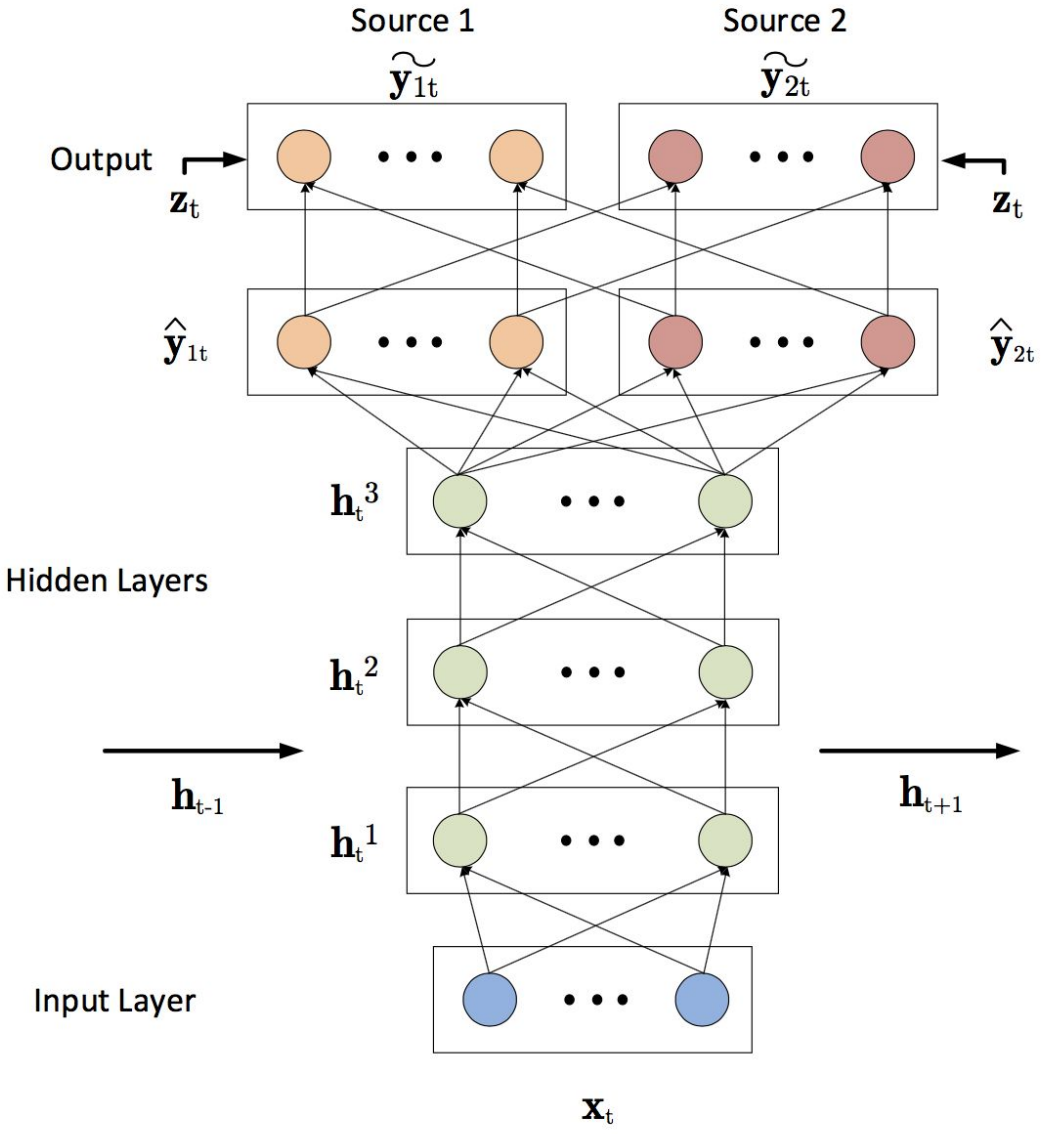
Bc. Lukáš Radoský

Akademický rok: 2019/2020

Úvod

Cieľom projektu je vytvoriť hlbokú neurónovú sieť schopnú separovať hlas interpreta alebo hudobný sprievod z pôvodných (zmiešaných) nahrávok. Jej vstupom je jednokanálový (mono) hudobný klip vo formáte *.wav*, pričom na výstupe je jednokanálový *.wav* súbor obsahujúci len hudobný podklad (karaoke) alebo interpretove vokály (a cappella).

Základný baseline projektu je postavený na existujúcej implementácii neurónovej siete podobného zamerania[1] [2] (rekurentné siete), ktorá separuje hlas aj hudbu z hudobného klipu súčasne, vetvením skrytých vrstiev pri konci sekvencie. Trénovanie jej autori uskutočňujú nad datasetom *MIR-1K*[2], ktorý obsahuje 110 karaoke piesní naspievaných prevažne amatérskymi participantmi.



Obrázok č. 1: Architektúra treťostrannej implementácie na báze RNN [2]

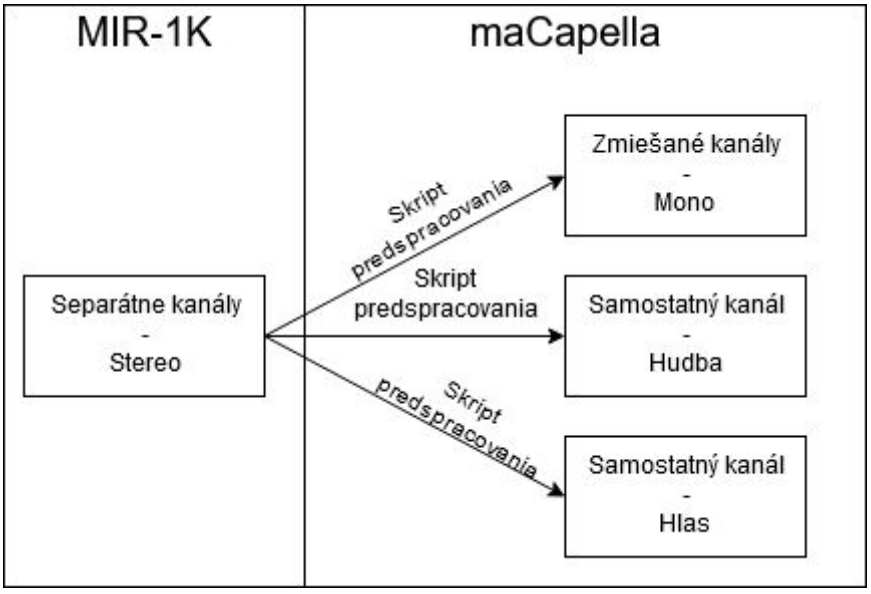
[1] https://github.com/andabi/music-source-separation
[2] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

Predspracovanie dát

Dataset *MIR-1K* využíva aj projekt *maCaPeLa*. Uloženie hudby a hlasu v separátnych kanáloch je veľmi výhodné v kontexte tejto problematiky. Predspracovanie dát prebieha v dvoch fázach:

- jednorazové predspracovanie,
- predspracovanie pri spustení (on-demand).

Jednorazové predspracovanie je vykonané len raz, mimo spustenia trénovania či validácie modelu. Daný dataset poskytuje hudobné klipy obsahujúce hudbu a hlas v separátnych kanáloch. Každý z klipov je k dispozícii aj rozdelený do viacerých menších klipov, pre pohodlnejšie spracovanie. Pre projekt *maCaPeLa* je potrebné pre každý z nich vytvoriť kópiu so zmiešanými kanálmi, ako aj kópiu s izolovaným kanálom (hudbou alebo hlasom, v závislosti od cieľa projektu). Tvorba týchto kópií je realizovaná v samostatných skriptoch, ktoré sú jednorazovo vykonané pred spustením modelu. Každé spustenie modelu potom operuje nad takto predspracovaným datasetom. Výhodou je zníženie výpočtovej náročnosti trénovania aj testovania.



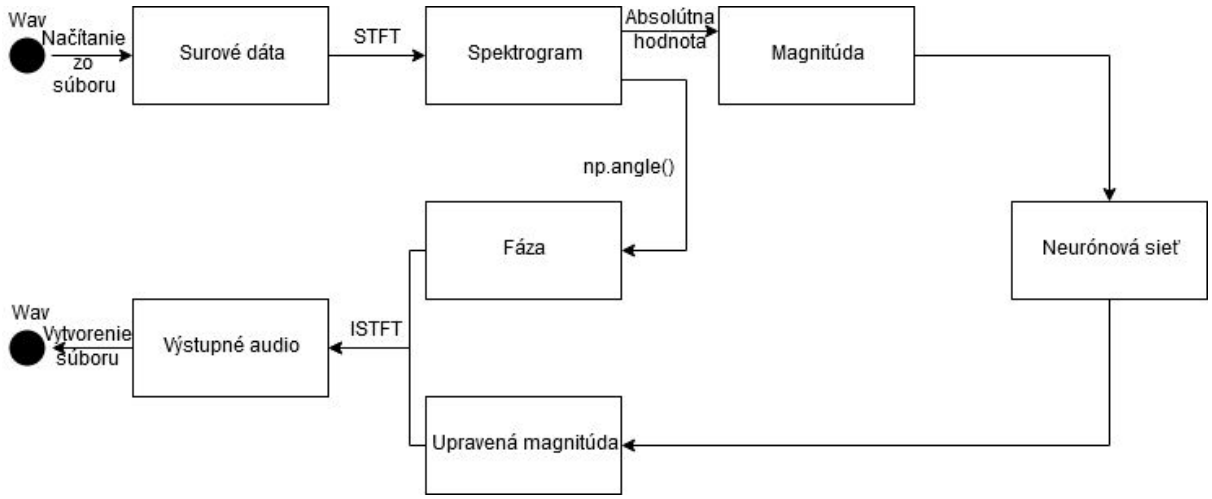
Obrázok č. 2: Jednorazové predspracovanie

Z pohľadu súčasného riešenia je podstatné vytvorenie súboru so zmiešanými kanálmi a so samostatným kanálom s hudbou (karaoke). Na základe jedného atribútu v konfiguračnom súbore je však možné ľahko zmeniť cieľ projektu na opačný - izoláciu hlasu a odstránenie hudby (a cappella).

Predspracovanie pri spustení je vykonané pri každom trénovaní modelu. Prvým krokom je načítanie audia zo súborov. Tým získavame surové (*raw*) dáta zvuku. V tejto podobe však manipulácia so zvukom nemá zmysel. Je potrebné vyoknať jeho transformáciu do spektrálnej (frekvenčnej) domény, teda získať jeho spektrogram. Toto je vykonané pomocou STFT (angl. short-time Fourier transform) vypočítanej knižnicou *librosa*. Spektrogram je následne prevedený na magnitúdu a fázu, kedy je zvuk rozdelený na úseky o zvolenej veľkosti okna, v našom prípade 1024. V tejto podobe je zvuk pripravený na spracovanie neurónovou sieťou. Dimenzionalita vstupu neurónovej siete by sa mala

zhodovať s dimenzionalitou jej výstupu. V prípade, že posledná dimenzia výstupu je väčšia ako 1 (vďaka poslednej vrstve), je nad touto dimenziou vykonaná agregácia, napr. priemer alebo medián.

Tok dát



Obrázok č.3: Tok dát v projekte

Pri načítavaní dát zo súboru vznikajú dvojice - audio obsahujúce pôvodnú pieseň a audio obsahujúce len cieľ separácie (hudba alebo hlas). V rámci predspracovania sú oba súbory transformované na magnitúdu a (bokom odloženú) fázu, čím vznikajú uniformne veľké úseky audia. Magnitúda týchto úsekov je vstupom pre neurónovú sieť, pričom:

- **x** je audio obsahujúce kanál so zmiešaným hlasom a hudbou, z ktorého má neurónová sieť izolovať požadovanú zložku.
- **y** je audio obsahujúce len zložku, izolovanie ktorej je úlohou neurónovej siete. Je očakávaným výstupom neurónovej siete.

V prípade, že sa vyžaduje aj vytvorenie výstupného súboru na základe predikcie neurónovej siete, je zo spektrogramu izolovaná aj fáza. Tá je skombinovaná s predikovanou magnitúdou pomocou inverznej STFT. Výsledkom sú surové dáta zvuku, z ktorých je následne vytvorený prehrateľný .wav súbor.

Architektúra modelu a experimentovanie

Model neurónovej siete je v rámci finálnej verzie postavený na konvolučných vrstvách. Jedná sa o 1D konvolúciu, pričom vstupné vektory majú dimenzionalitu [1024, 1]. Pri experimentoch sme využili 1D konvolučné vrstvy, max pooling vrstvy, plne prepojené vstvy a flatten vrstvy, s rôznymi počtami neurónov či aktivačnými funkciami. Finálnu architektúru, teda tú s najlepšími výsledkami uvádzame až po opise experimentov.

Pri experimentoch sme sa pokúšali zostaviť rôzne architektúry pomocou uvádzaných vrstiev. Tiež sme experimentovali s počtom neurónov a aktivačnými funkciami. Z globálnejšieho pohľadu sme experimentovali s rôznymi hyperparametrami vrátane počtu vzoriek, pomeru veľkosti trénovacej a testovacej množiny, loss funkcie, počtu epoch či veľkosti dávok (angl. batch count).

Kvalitu modelov sme overovali dvojako:

- **objektívne** - pomocou číselných metrík, akými sú loss funkcia, presnosť na trénovacej vzorke a presnosť na testovacej vzorke;
- **subjektívne** - pomocou zvoleného hudobného súboru, pre ktorý sme po ukončení trénovania daného modelu urobili predikciu a vypočuli si ho. Číselné metriky totiž môžu ukazovať na dobré výsledky, ale reálne audio v niektorých prípadoch môže dosahovať nedostatočnú kvalitu v rôznych ohľadoch (najčastejšie degradácia “farby” a dynamiky hlasu).

Počas experimentov sa ukázalo, že navrhované architektúry a konfigurácie vykazujú omnoho lepšie výsledky pri izolovaní hlasu než pri izolovaní hudby. Preto sme sa primárne **zamerali na problém izolovania interpretovho hlasu (a cappella)**.

Pri experimentovaní sme skúšali viaceré možné kombinácie architektúr a parametrov, pričom sme sa snažili o sledovanie charakteristík, ktoré sme považovali za výrazne ovplyvňujúce kvalitu výsledkov. Nižšie popisujeme vykonané experimenty formou tabuľky, kde hrubá horizontálna čiara oddeľuje jednotlivé konfigurácie modelu.. Viaceré kombinácie pre prehľadnosť neuvádzame, nakoľko sme sa pomocou nich neposunuli k lepším výsledkom (slepé cesty experimentovania).

Vrstvy	Parametre vrstvy		Počet vzoriek	Test. vzorka	Epochy	Acc.	Val. acc.	Subj. val.	Označenie
	Parameter	Hodnota							
Conv1D	units	50	10	50%	3	24%	15%	Množstvo šumu, umelo znejúci hlas, podprieme rné výsledky separovani a.	training_06_12_2019_18_25_03
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Flatten	N/A	N/A							
Dense	units	1024							
	activation	relu							
Conv1D	units	50	30	50%	5	95%	93%	Lepšie potlačenie hudby, avšak výrazné zhoršenie kvality celkovej nahrávky	training_06_12_2019_19_11_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50	40	50%	5	47%	42%	Analogicky vyššie, mierne lepšia kvalita.	training_06_12_2019_19_23_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50	40	50%	5	76%	69%	Priemerné potlačenie	training_06_12

	kernel_size	10						hudby, zhoršenie celej nahrávky (pod vodou)	_2019_19_30_22
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50	40	50%	5	8%	6%	Mierne potlačenie hudby, výrazné zhoršenie celej nahrávky.	training_06_12_2019_19_35_22
	kernel_size	10							
	padding	same							
	activation	relu							
Conv1D	units	50	40	50%	5	90%	89%	Dobré potlačenie hudby, akceptovateľná kvalita výsledku.	training_06_12_2019_19_46_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50	40	50%	5	50%	59%	Veľmi dobré potlačenie hudby; mierne umelý	training_06_12_2019_19_54_22
	kernel_size	10							
	padding	same							

	activation	relu						dojem; oscilácia hlasitosti.	
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50	40	50%	5	65%	64%	Analogicky vyššie.	training_06_12_2019_19_57_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							

	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	15	40	50%	5	51%	51%	Priemerné potlačenie hudby, zhoršenie celej nahrávky.	training_06_12_2019_20_05_22
	kernel_size	10							
	padding	same							
	activation	relu							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
Conv1D	units	45							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2	40	50%	5	1%	1%	Mierne potlačenie hudby, katastrofálna kvalita.	training_06_12_2019_20_09_22
	strides	1							
	padding	same							
MaxPooling1D	pool_size	2							
	strides	1							

	padding	same							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	15	40	50%	5	42%	45%	Dobré potlačenie hudby; silno umelý dojem; oscilácia hlasitosti.	N/A
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	45							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	45							
	kernel_size	10							
	padding	same							

	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	45							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	15	40	20%	10	31%	35%	Priemerné potlačenie hudby; šum; zhoršenie kvality nahrávky.	N/A
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	60							
	kernel_size	10							
	padding	same							

	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Dense	units	1024							
	activation	relu							
Conv1D	units	15	40	20%	30	41%	41%	Analogicky vyššie; mierne vyššia kvalita.	N/A
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	60							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Dense	units	1024							

	activation	relu							
Conv1D	units	15	40	20%	15	36%	41%	Veľmi dobré potlačenie hudby; pomerne umelý dojem z výslednej nahrávky.	training_07_12_2019_11_49_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	60							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	60							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							

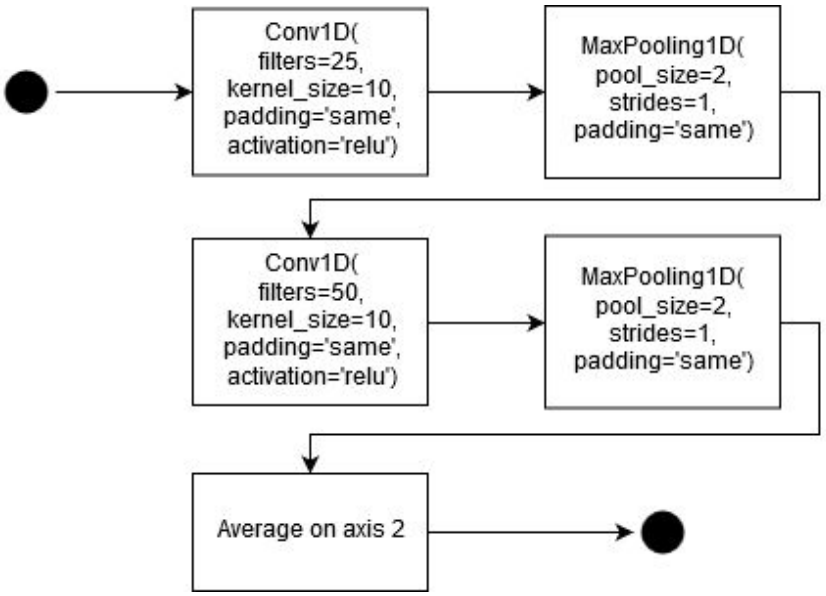
	padding	same							
Conv1D	units	60							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	25	100	20%	15	41%	60%	Priemerné potlačenie hudby; šum; zhoršenie kvality nahrávky.	training_07_12_2019_12_11_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Dense	units	1024							
	activation	relu							
Conv1D	units	25	100	20%	10	52%	50%	Dobré potlačenie hudby, zachovaná integrita a kvalita	training_07_12_2019_12_58_55
	kernel_size	10							
	padding	same							
	activation	relu							

MaxPooling1D	pool_size	2						hlasu; slabšie odfiltrovanie basovej zložky.	
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	15	100	20%	10	68%	79%	Dobré potlačenie hudby, mierne umelý dojem.	training_07_12_2019_13_19_22
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	30							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	45							
	kernel_size	10							
	padding	same							
	activation	relu							

MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	25	300	20%	12	57%	56%	Dobré potlačenie hudby, zachovaná integrita a kvalita hlasu	training_07_12_2019_13_54_32
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	25	300	20%	12	46%	41%	Analogicky prípadom vyššie.	training_07_12_2019_14_38_53
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							

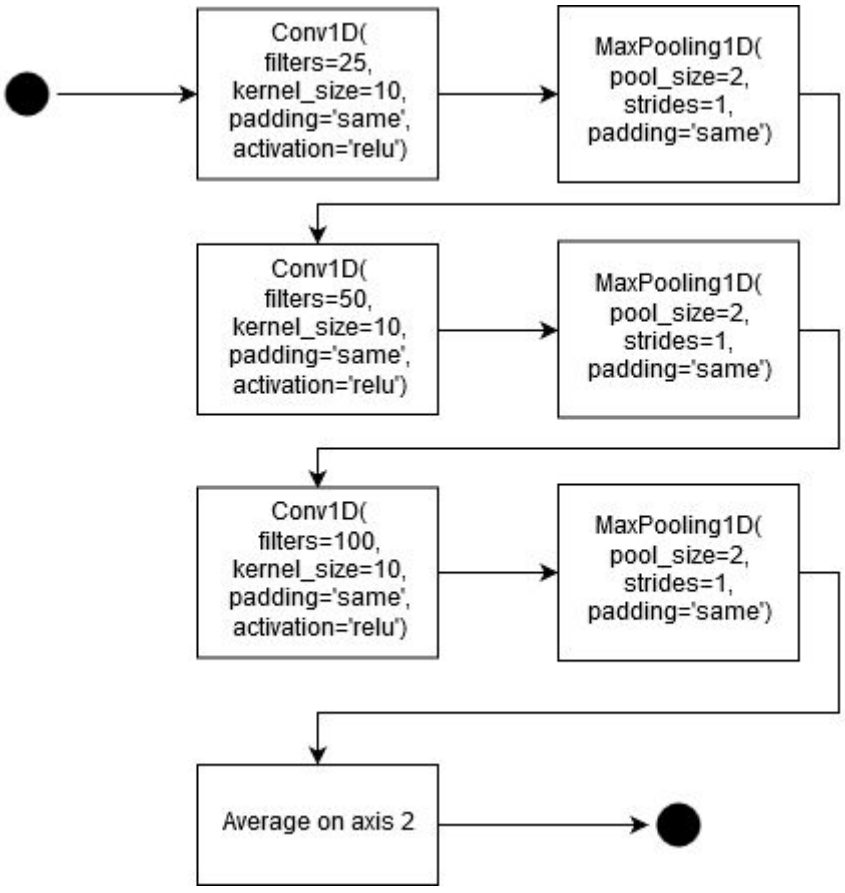
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	25	200	20%	12	45%	51%	Veľmi dobré potlačenie hudby /i basov); rozumná miera zachovania integrity hlasu.	training_07_12_2019_15_20_34
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	50							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							
Conv1D	units	100							
	kernel_size	10							
	padding	same							
	activation	relu							
MaxPooling1D	pool_size	2							
	strides	1							
	padding	same							

Ako najlepšie sme zhodnotili dve vybrané architektúry (zelené zvýraznenie), najmä podľa subjektívnej metriky - vypočutie si nahrávok.



Obrázok č.4: Najlepšie architektúry - behy training_07_12_2019_12_58_55 a training_07_12_2019_13_54_32

Táto architektúra vykázala výborné výsledky až v dvoch behoch, s drobnou odlišnosťou v hyperparametroch. V tomto prípade sa jedná o prekvapivé výsledky, nakoľko ide o pomerne jednoduchú architektúru. Hoci objektívne metriky v týchto prípadoch nevykazujú príliš dobré hodnoty (presnosť v rozmedzí 50-60%), audio vytvorené takto natrénovaným modelom malo výbornú kvalitu. Ukazuje sa tak, že vybraná úloha je veľmi náročná na objektívne meranie a posúdenie. Subjektívne sa zdá byť najlepšou nahrávkou tá, ktorá vznikla pri behu *training_07_12_2019_13_54_32*.



Obrázok č.5: Najlepšie architektúry - beh training_07_12_2019_15_20_34

Veľmi dobré výsledky tiež vykázala podobná architektúra, rozšírená o jednu dvojicu konvolučnej a pooling vrstvy, so zdvojnásobeným počtom neurónov v porovnaní s predošlou. Presnosť v tomto prípade tiež pôsobí neveľmi úspešne (cca 40%-50%), avšak výstupné audio vykazuje doposiaľ najlepšie potlačenie hudobnej zložky bez poškodenia hlasovej zložky. Nepresahuje však kvalitu výstupu architektúry z obrázku č. 4.

Zhodnotenie

V rámci tohto semestrálneho projektu bolo našou úlohou natrénovať hlbokú neurónovú sieť pre účely riešenia úlohy z oblasti spracovania audia - automatizovaná tvorba a cappella (primárne zameranie práce), resp. karaoke verzií piesní z originálnych nahrávok, obsahujúcich hlasovú zložku (vokály), ako aj hudobný sprievod. Pre tieto účely sme v rámci viacerých iterácií implementovali funkčný model na báze konvolučnej neurónovej siete (naše prvotné experimenty skúmali možnosť využitia rekurentných sietí, ktoré sa však ukázali byť pre naše účely nevhodné), pričom sme v kontexte evaluácie jednotlivých verzií modelu vyhodnocovali stanovené objektívne (mean squared error) i subjektívne metriky (kvalita výslednej nahrávky na základe jej vypočutia).

Najlepšie výsledky sme dosiahli s modelmi, obsahujúcimi 2, resp. 3 dvojice vrstiev v konfigurácii (konvolučná vrstva, MaxPooling vrstva) bez použitia dodatočných typov vrstiev (Dense, resp. Flatten). Zatiaľ čo model s 3 špecifikovanými dvojicami vrstiev dokázal oddeliť hlasovú zložku od nahrávky v najväčšej možnej miere, za pozornosť stojí tiež model s 2

dvojicami, ktorý dokázal lepšie zachovať pôvodnú integritu i dynamický rozsah hlasu za cenu mierne vyššieho zastúpenia (neželanej) basovej zložky vo výsledku.

Ako možné ďalšie smerovanie projektu možno navrhnúť využitie kvalitnejšieho datasetu (vyššia vzorkovacia frekvencia; v našom prípade iba 16 kHz), trénovanie na väčšom množstvo vzoriek v kombinácii s využitím výkonnejšieho výpočtového hardvéru (grafické karty namiesto procesora), ako aj dodatočný pre-, resp. post-processing nahrávok (normalizácia úrovní či využitie techniky časovo-frekvenčného maskovania [1,2] -- výstupom siete by v tomto prípade nebola finálna nahrávka, iba filter, ktorý možno následne aplikovať na pôvodný vstup).

Bibliografia

[1] CHANDNA, P. et al. Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. In: *International Conference on Latent Variable Analysis and Signal Separation* [online]. Barcelona: Universitat Pompeu Fabra, 2017. Dostupné na internete: http://mtg.upf.edu/node/3680.

[2] HUANG, P. et al. Singing-voice Separation from monaural recordings. In: *International Society for Music Information Retrieval Conference (ISMIR)* [online]. Illinois: University of Illinois, 2014. Dostupné na internete: https://posenhuang.github.io/papers/DRNN_ISMIR2014.pdf%0D.