Klasyfikacja gatunków muzycznych za pomocą sieci neuronowych

Rafał Bojarczuk, Igor Szołucha, Zhanna Solobchuk, Teodor Stupnicki, Piotr Sosno

16 06 2020

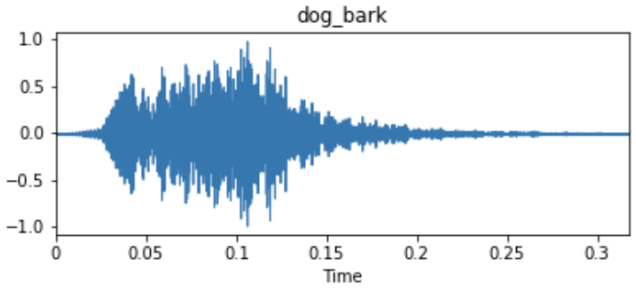
1. Wprowadzenie

Celem naszego projektu jest przypisanie utworu muzycznego do jednego z pięciu gatunków: muzyka klasyczna, muzyka elektroniczna, rock, pop oraz kip-hop.

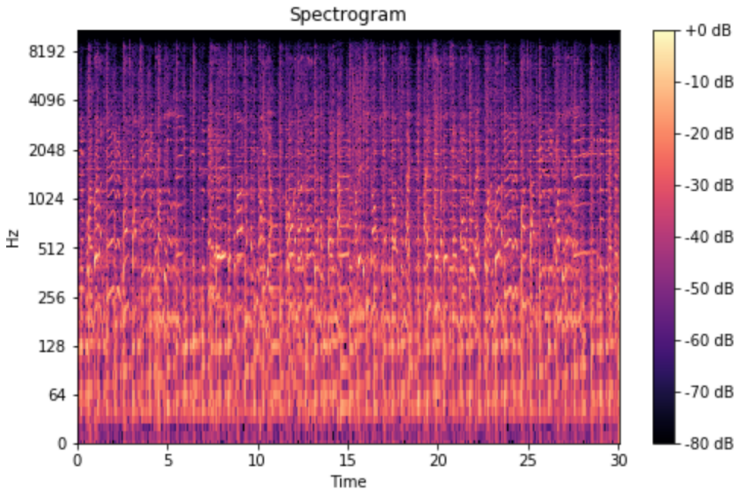
Aby rozwiązać nasz problem musieliśmy odpowiedzieć sobie na pytanie, jaki problem tak na prawdę rozwiązujemy. Jest to oczywiście problem klasyfikacji, jeden z najbardziej znanych i codziennie spotykanych problemów. Jest to zadanie klasyfikacji wieloklasowej - w której klasy stanowią wymienione wyżej gatunki muzyczne.

1.1 Podejście

Naszymi danymi testowymi są krótkie fragmenty utworów o znanym nam gatunku, i to właśnie z tych fragmentów musimy zebrać wszystkie potrzebne nam dane do przeprowadzenia dalszej analizy. Najczęściej spotykaną formą przedstawienia tych próbek jest szereg czasowy obrazujący amplitudę fal dźwiękowych.



Jednak z takiego wykresu odczytanie znaczących informacji jest prawie niemożliwe, dla różnych gatunków muzycznych wykresy były do siebie bardzo zbliżone. W związku z tym zdecydowaliśmy się użyć spektogramów. Jest to graficzna reprezentacja natężenia sygnału o różnych częstotliwościach w czasie, a wyznaczyć ją można za pomocą Krótkotrwałej transformaty Fouriera (short-time Fourier transform) obliczanej dla kolejnych okien czasowych.



W naszym projekcie zdecydowaliśmy się na użycie sieci neuronowych. Był to dość oczywisty wybór ze względu na złożoność naszego problemu, typ danych wejściowych i ich wielowymiarowość. Również związek otrzymanych ze spektogramów danych z gatunkiem muzyki nie zawsze był dla nas jasny, a w takich sytuacjach sieci neuronowe wydają się jedyną rozsądną opcją. Jako, że spektrogramy to dane obrazowe i wartości pikseli znajdujących się blisko siebie są ze sobą blisko powiązane (reprezentują dźwięki o podobnych częstotliwościach lub następujące po sobie) to do rozwiązania problemu użyjemy w sieci warstw splotowych (convolutional layers).

2. Zbiór danych

Przy dobraniu właściwego zbioru danych zastanawialiśmy się nad poniższymi zbiorami:

- GTZAN składa się z 1000 ścieżek audio po 30 sekund. Zawiera 10 gatunków, a mianowicie blues, klasyczna, country, disco, hip hop, jazz, reggae, rock, metal i pop. Zbiór zawiera 100 ścieżek każdego gatunku.
- FMA (wersja large) zawiera 106574 utworów po 30 sekund. z 16341 artystów i 14854 albumów i 161 niezbalansowanych gatunków. Zbiór FMA full zawiera taką samą liczbę utworów, ale o pełnym rozmiarze a nie 30 sekundowe kawałki.
- MilionSongsDataset zawiera 1 000 000 plików które zawierają featury wielu nutek, a nie pliki audio. Przy pracy z CNN woleliśmy mieć pliki audio więc zrezygnowaliśmy z tego zbioru.

Najważniejszymi kryteriami wyboru zbioru danych były liczba utworów oraz dostępność interesujących nas gatunków. Po analizie zbiorów doszliśmy do wniosku, że najlepszym rozwiązaniem skorzystanie z połączenia zbiorów GTZAN i FMA w wersji large. Zbiory train/validation/test są utworzone według domyślnego podziału zaproponowanego przez twórców zbioru FMA o proporcjach około 80/10/10. Zbiór GTZAN jest podzielony według dokładnie takich proporcji i odpowiadające sobie części zostały ze sobą połaczone, tak aby zbiory treningowy, walidacyjny i testowy pochodziły z podobnych rozkładów.

```
In [35]: large[('track', 'genre_top')].value_counts()

Out[35]: Rock                  14182
         Experimental          10608
         Electronic             9372
         Hip-Hop                3552
         Folk                   2803
         Pop                    2332
         Instrumental           2079
         International          1389
         Classical              1230
         Jazz                    571
         Old-Time / Historic     554
         Spoken                  423
         Country                 194
         Soul-RnB                175
         Blues                   110
         Easy Listening           24
         Name: (track, genre_top), dtype: int64
```

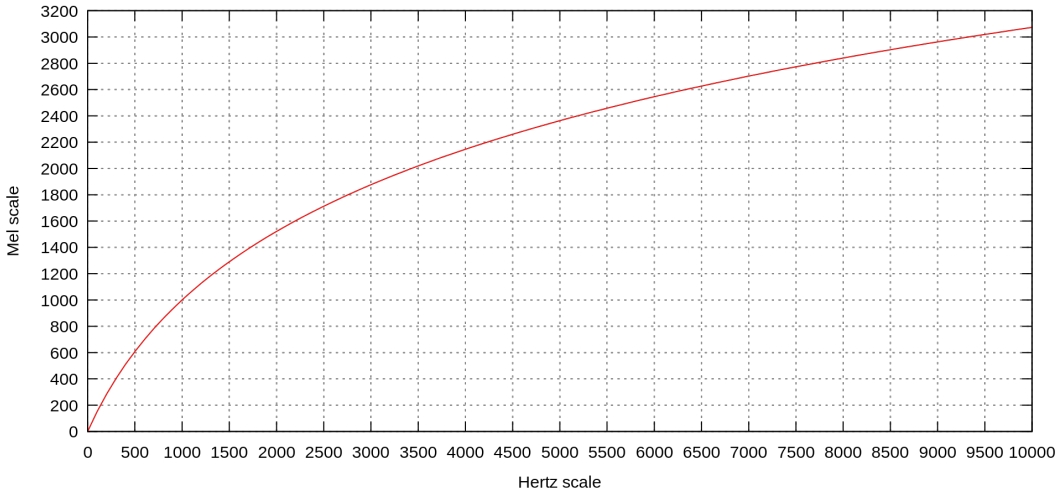
Żeby uniknąć dysproporcji spowodowanej dużą ilością Rock i Electronic muzyki postanowiliśmy dodać tylko połowę reprezentantów tych klas ze zbioru FMA.

Po usunięciu plików, które okazały się wadliwe, zostaliśmy z 19147 30 sekundowymi utworami

Gatunek	FMA	GTZAN
klasyczna	1226	100
elektroniczna	4587	0
rock	7069	100
pop	2328	100
hip hop	3537	100
Łącznie	18747	400

3. Przygotowanie danych - tworzenie spektrogramu

Aby uzyskać dane wejściowe do sieci tworzymy spektrogramy w skali Mel. Jest to skala logarytmiczna, która dostosowuje różnice w częstotliwości dźwięków do czułości ludzkiego ucha. Ludzki mózg postrzega dwa dźwięki o częstotliwościach 500Hz i 1000Hz jako całkiem różne, a dźwięki o wysokości 8000Hz i 8500Hz jako prawie takie same, mimo że bezwzględna różnica na skali w Hertzach jest taka sama.



Również natężenie dźwięku konwertujemy do skali logarytmicznej i wyrażamy w decybelach.

Często stosowanym zabiegiem jest dodatkowe obliczenie DCT - dyskretnej transformaty kosinusowej (Discrete Cosine Transform), która pomaga w ekstrakcji i dekorelacji cech. Otrzymujemy wtedy MFCC - Mel-frequency cepstral coefficients, współczynniki które opisują spektrogram w bardziej skompresowanej formie, lecz w przypadku korzystania z sieci neuronowych sam Mel spektrogram zazwyczaj daje nieco lepsze efekty, jako że przy dużej ilości danych sieć neuronowa dobrze radzi sobie z promowaniem najbardziej różnicujących cech.

Dysponujemy 30 sekundowymi fragmentami audio z częstością próbkowania 22050Hz. Długość skoku szybkiej transformaty Fouriera ustaliliśmy jako 2048 próbek, co daje 322 (30*22050/2048) okna w których obliczona jest transformata, a więc jako niezmienny moment w muzyce przyjmowane jest niecałe 0,1 sekundy. Długość okna wynosi 4096 próbek, co sprawia, że sąsiadujące okna na siebie nachodzą, aby nie stracić żadnej informacji pomiędzy nimi. Liczba cech, czyli różnych częstotliwości w skali mel wynosi 128 dla każdego okna czasowego. To daje nam spektrogram o wymiarach 128x322, ale nie każdy plik ma długość dokładnie 30 sekund, więc ucięliśmy wymiary do 128x320 aby wyrównać je z nieco krótszymi ścieżkami.

4. Architektura sieci

Konwolucyjne sieci neuronowe CNN to jeden z wariantów sieci neuronowych często wykorzystywanych w dziedzinie wizji komputerowej. Nazwa pochodzi od rodzaju ukrytych warstw, z których się składa. Sieci konwolucyjne zwykle składają się z trzech typowych części:

1. Warstwa konwolucyjna zawierająca zbiór adaptowalnych filtrów (Convolutional layer)

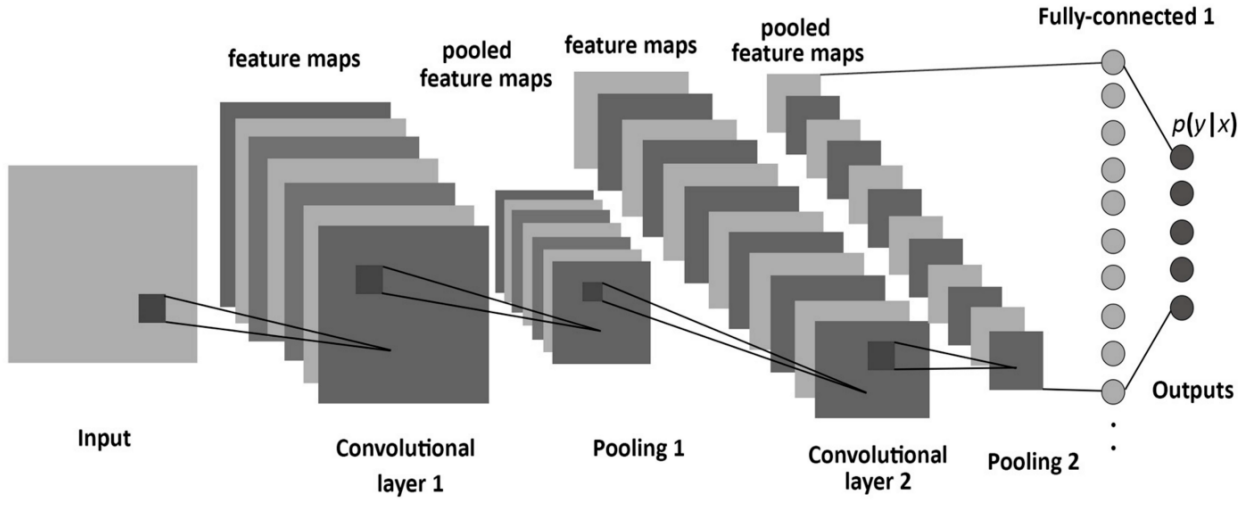
Konwulacja działa na dwóch sygnałach/obrazach, przy czym jeden z nich jest sygnałem wejściowym a drugi filtrem. Celem tej warstwy jest zwielokrotnienie sygnału wejściowego z jądrem, aby uzyskać zmodyfikowany sygnał.

2. Warstwa łączenia (Pooling layer)

Proces łączenia polega na dyskretyzacji i zagregowaniu otrzymanych danych. Celem jest zmniejszenie wymiarów badanej próbki.

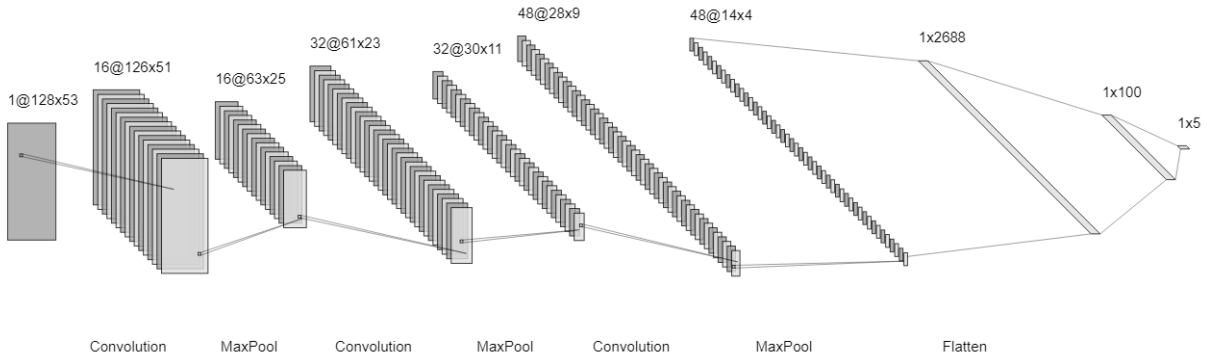
3. Warstwa połączeń każdy z każdym

Po zastosowaniu wybranej ilości warstw konwolucyjnych i łączących otrzymana macierz zostaje spłaszczona i stanowi wejście do standardowej sieci neuronowej, stworzonej z w pełni połączonych warstw.



4.1 Nasza architektura

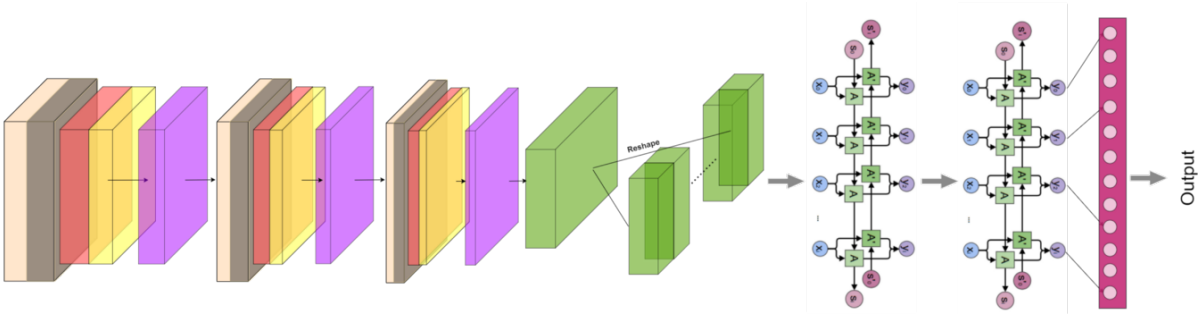
Najlepsze wyniki udało nam się osiągnąć na sieci o następującym schemacie:



Wykorzystalimy 3 sekwencje następujących po sobie warstw; warstwy konwolucyjnej, warstwy łączącej i operacji dropout z prawdopodobieństwem 20%. Po przejściu przez ten etap dane były transformowane do jednego wymiaru, a następnie zagęszczane. Najpierw zmieniliśmy wymiar do 100, a w następnej operacji do ilości rozpoznawanych przez kategorii. Jako funkcji aktywacji używaliśmy ReLU, poza ostatnią warstwą odpowiadającą za klasyfikację, gdzie funkcją aktywacji była funkcja softmax. Na schemacie można zauważyć, że dane wejściowe są rozmiarów 128x53, w następnych rozdziałach się to wyjaśni.

4.2 Inne podejścia

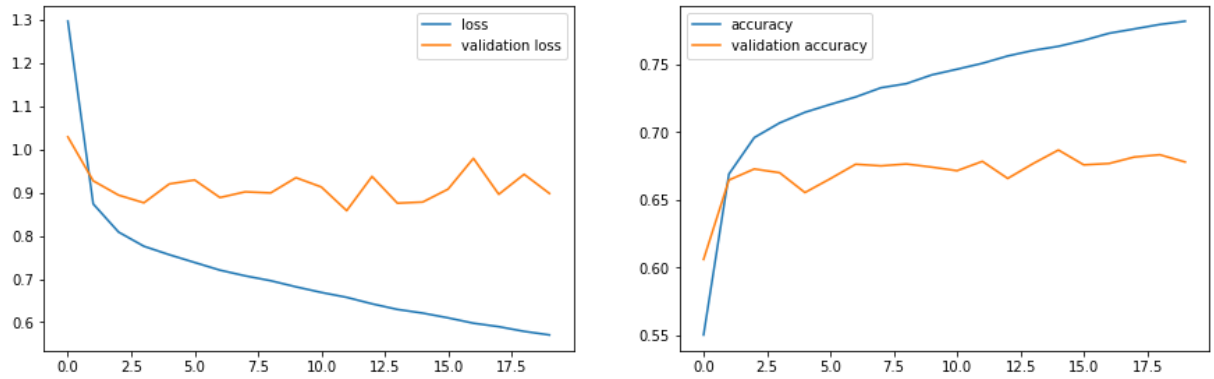
Konwolucyjno-rekurencyjne sieci neuronowe CRNN jest połączeniem zwykłych sieci konwolucyjnych z sieciami rekurencyjnymi. Kombinacja obu rodzajów pozwala na wykorzystanie konwolucyjnych sieci do wyodrębnienia cech z próbki w danym momencie w czasie, a następnie przekazanie następujących po sobie próbek do warstwy rekurencyjnej, odpowiedniej do analizy danych sekwencyjnych.



W naszym przypadku, mimo obiecującej teorii nie udało się uzyskać lepszego rezultatu architekturą konwolucyjno-rekurencyjną. Podczas testów zetknęliśmy się z dużym stopniem nadmiernego dopasowania (overfitting), którego nie udało się nam wyeliminować zmieniając parametry sieci. Powodem najprawdopodobniej jest zbyt mała ilość danych w stosunku do ich złożoności, być może właśnie dlatego, odpowiednie wytrenowanie bardziej skomplikowanej sieci się nie udało.

5. Zwiększenie zbioru

Jak wspomnieliśmy wcześniej, głównym problem z którym się mierzyliśmy był overfitting. Od pewnego momentu celność sprawdzana na danych walidacyjnych wachała się na poziomie 60-65% i nie rosła już mimo dalszego treningu.



Zdecydowaliśmy się sztucznie zwiększyć zbiór danych poprzez podzielenie spektrogramów zawierających dane z 30 sekund na mniejsze części. Zazwyczaj człowiekowi wystarczy kilka do kilkunastu sekund, żeby móc rozpoznać gatunek słuchanej muzyki, uznaliśmy że sieć neuronowa również sobie poradzi. Najbardziej efektywnym okazał się podział na 6 równych części po 5 sekund. Po podzieleniu otrzymaliśmy 90192 spektrogramów w zbiorze trenującym, 10854 w validation set oraz 10350 w zbiorze testowym. Po tym zabiegu dane wejściowe zmieniły rozmiar do 128x53, co widać na schemacie w rozdziale 4. oraz celność architektury zwiększyła się o kilka punktów procentowych.

6. Analiza wyników

Korzystając z funkcji classification_report() narzędzi sklearn pozyskaliśmy informacje o jakości naszej klasyfikacji.

	precision	recall	f1.score
Classical	0.8643411	0.7729636	0.8161025
Electronic	0.7058337	0.6968954	0.7013361
Pop	0.3989770	0.1235154	0.1886336
Rock	0.7488925	0.9097358	0.8215154
Hip-Hop	0.7752427	0.8090172	0.7917700
accuracy	0.7366184	0.7366184	0.7366184
macro avg	0.6986574	0.6624255	0.6638715
weighted avg	0.7074701	0.7366184	0.7098855

Precision Precyzja jest zdefiniowana jako iloraz obserwacji przypisanych do prawidłowej klasy i wszystkich obserwacji przypisanych do tej klasy. Pozwala nam ona określić jak wiarygodne są nasze wyniki

$$Precision = \frac{true\_positive}{true\_positive + false\_positive}$$

True Positive - Klasa prawdziwie pozytywna; False Positive - Klasa fałszywie pozytywna

Recall Miara recall jest zdefiniowana jako iloraz obserwacji przypisanych do prawidłowej klasy i wszystkich obserwacji, które powinny zostać przypisane do tej klasy. Pozwala nam ona określić jaką część wyników gubimy

$$Recall = \frac{true\_positive}{true\_positive + false\_negative}$$

False Negative - Klasa fałszywie negatywna

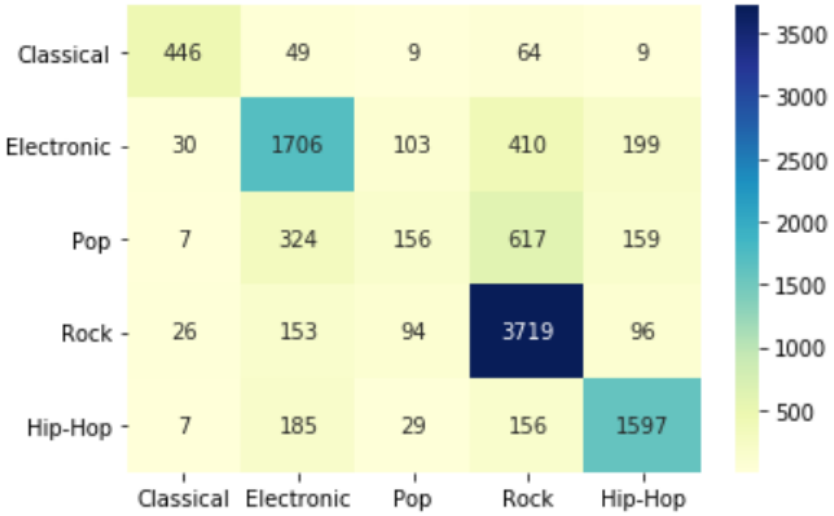
F1 Score Jest to miara dokładności klasyfikatora równa średniej harmonicznej z precision i recall. Obliczamy ją w następujący sposób

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

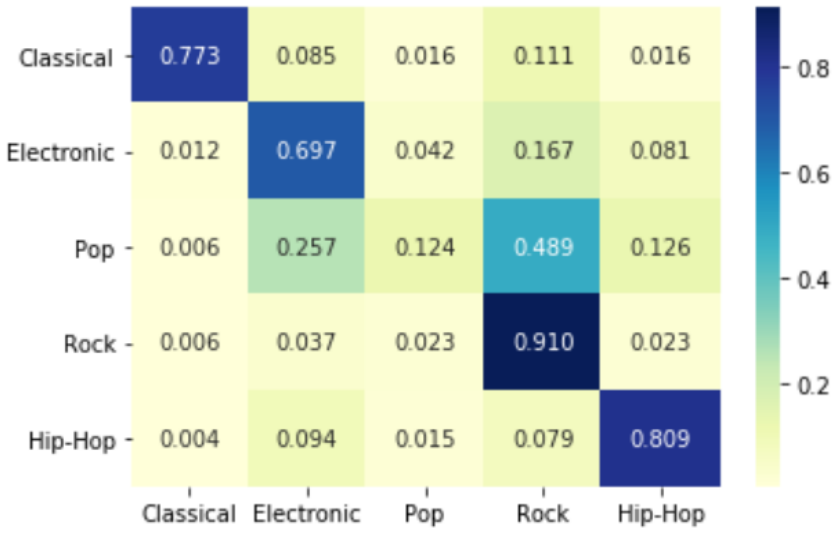
Najlepsza dokładność (accuracy) z jaką udało nam się sklasyfikować utwory to: 73,66%, korzystając z architektury CNN. W przypadku architektury CRNN najlepsza dokładność wynosiła około 66%. Zrezygnowaliśmy z rozwijania tej architektury ze względu na to, że dawała gorsze wyniki od zwykłych sieci konwolucyjnych.

Precyzja najlepiej prezentuje się dla gatunku Classical - 0.86, oprócz niego dla gatunów Electronic, Rock i Hip-Hop precyzja nieznacznie odbiega. Duża różnica w precyzji występuje w gatunku Pop, jest on słabo klasyfikowany. Dla tego gatunku, również wartość wielkości recall jest niezadowalająco niska. Największa wartość recall'u została odnotowana dla gatunku Rock - ponad 0.9, oznacza to, że ponad 9/10 utworów z tego gatunku zostało jako właśnie ten gatunek sklasyfikowane.

Macierz pomyłek



Macierz pomyłek pokazuje nam jak były klasyfikowane poszczególne gatunki muzyczne, to co rzuca się w oczy najbardziej, to fatalne wyniki dla muzyki popularnej. Dużo częściej było ona klasyfikowana jako Rock lub muzyka elektroniczna. Dzieje się tak dlatego, że jest to bardzo zróżnicowana kategoria. Gatunki muzyczne mogą się ze sobą łączyć i przeplatać, a muzyka popularna w swojej definicji nie ma ustalonych ram tego jak powinna brzmieć, często korzysta właśnie z tego, co w danym czasie jest popularne i miesza ze sobą najpopularniejsze i najszerzej pożądane brzmienia. Co ciekawe, kiedy pominiemy dane dla muzyki pop, pozostałe 4 gatunki są klasyfikowane z celnością wynoszącą 80,8%.

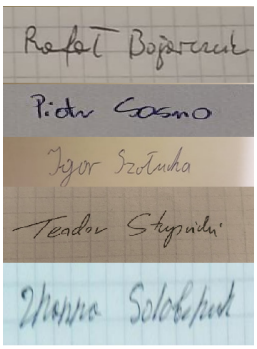


Powyżej przedstawiamy przeskalowaną macierz, mapa kolorów jest dostosowana proporcjonalnie do liczby utworów w każdym gatunku.

7. Podsumowanie

73,66% celności dla pięciu i 80,8% dla czterech gatunków muzycznych to całkiem zadowalający wynik dla prostej architektury. Z pewnością byłby wyższy, gdybyśmy mieli możliwość skorzystania z pełnej wersji zbioru fma, niestety szybkość łącza oraz rozmiary naszych dysków nie pozwoliły na pobranie 971 gigabajtów danych i musieliśmy się ograniczyć do zbioru zawierającego 30 sekundowe fragmenty utworów.

`Potwierdzam samodzielność powyższej pracy oraz niekorzystanie przeze mnie z niedozwolonych źródeł`



Źródła

https://en.wikipedia.org/wiki/Mel_scale

https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90

https://towardsdatascience.com/an-approach-towards-convolutional-recurrent-neural-networks-a2e6ce722b19

http://alexlenail.me/NN-SVG/LeNet.html?fbclid=IwAR2bSfVQ-vjhTr0EDtHq9HzeOTuiABARc-qCGS-O0Wnr34-ThQfXaNTLLdU