Madárhangfelismerés konvolúciós neurális hálózat segítségével (CNN)

Convolutional Convivality

Hegedüs László (G0J9RT)

Mihályi Márk (EAPH9W)

Kozák Balázs (DXAINN)

Absztrakt

A madarak a környezet egészségét tükrözik, mivel a szennyezés és az éghajlatváltozás hatással van az ökológiára [1].

A gépi tanulás szakértői rengeteg ismeretet dolgoznak fel a biológiai sokféleség folytonos léptékű nyomon követése során. Napjainkban a konvolúciós neurális hálózatok (CNN-ek) nagy előnnyel bírnak a különböző fajok azonosítására [2].

A konvolúciós neurális hálózatok (CNN-ek) erősek a gépi tanulás eszközkészletei, amelyek hatékonynak bizonyultaka képfeldolgozás és hangfelismerés területén [1].

Jelen cikkben a CNN rendszer a madárhangokat osztályozza különböző konfigurációkon és hiperparamétereken keresztül bemutatva és tesztelve [2].

A spektrogramok a generált, letöltött adatok neurális hálózat bemenetét jelentik. A mellékelt kísérletek különböző konfigurációkat hasonlítanak össze beleértve az osztályok számát (madárfajok) és a hangszínt beleértve.
Az eredmények azt sugallják, hogy a színkép (spektogram) kiválasztása a képekkel összhangban van. Az előre tanult hálózattal lényeges idő spórolható meg [2].

Short-Time Fourier Transform (STFT) és Mel Frequency Cepstral együttható (MFCC) algoritmusokat használnak a spektrogramképek elkészítésére [1].

A Mel-Frequency Cepstral Coefficients (MFCC) egy nagyon népszerű módszer a beszédfelismerésben, hasonló technika alapján működik a modell, mint az emberi érzékelés [3]. Valójában ezt modellezi.

A hanganyagok különböző hosszúságúak (időtartománybeli) a tanulmány elkészítéséhez azonos hosszúságú hanganyokra van szükség. A madárhangokat spektrogramképekké alakítva, a frekvenciák spektruma (függőleges y-tengely, Hz) idő szerint (vízszintes x-tengely, mp). Az egyes pixelek intenzitása a hang amplitúdóját jelenti[4].

Gyors Fourier transzformáció (FFT), Mel Filter Bank, diszkrét koszinusz transzformáció plusz delta jellemző, ezen értékek együttes kimenete az MFCC együttható [5].

Bevezetés

A hangfájlok könnyebben előállíthatóak a gyakoribb fajok esetében, míg ritkább fajoknál az adatbővítés (adatdúsítást) szinte rendszeresen alkalmazzák. Az adatok elérhetősége nagy kihívás CNN-eket alkalmazva, másnéven az adatok begyűjtése, előkészítése, szegmentálása [6].

A frekvencia tartalmat reprezentáló spektrogramképek a hangszínek. Két algoritmus, rövid idejű Fourier Transform (STFT) és Mel Frequency Cepstral, (MFCC) a hang konvertálására szolgálnak [6,7].

Az STFT-t úgy alkalmazzák az audiojelre, hogy a jelet felosztják külön átfedő keretekre, majd kiszámítják a DTFT-t minden egyes képkockához ami egy komplex mmátrixot eredményez [7].

$$STFT\{X\}(m,\omega) = Xm(\omega) = \sum_{n=-\infty}^{\infty} x_n \omega(n - mR) e^{-j\omega n}$$

1. ábra Rövid idejű Fourier transzformáció

ahol Xm a bemeneti jel az n időpontban, az ω(n) Hann ablak, amelynek hosszúsága m=1024, középpontja n, és R = 256 a ugrásméret az egymást követő képkockák között. A Hann nagy ablaka 1024-es 75%-os átfedéssel rendelkezik. Az STFT kiszámítása után lehet [7].

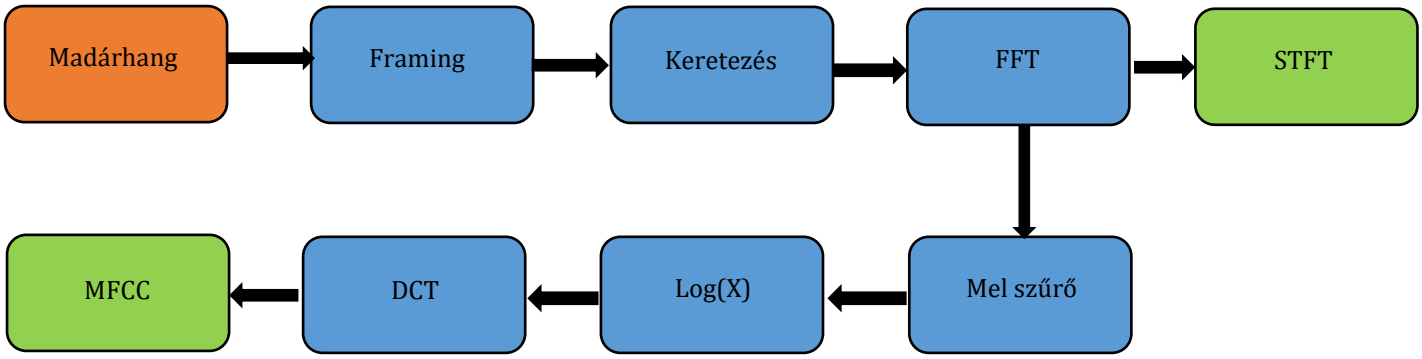
Az MFCC kiszámításához használt egyenlet szerint.

$$Mel\ (f) = 2595\ log10\ (1 + \frac{f}{700})$$

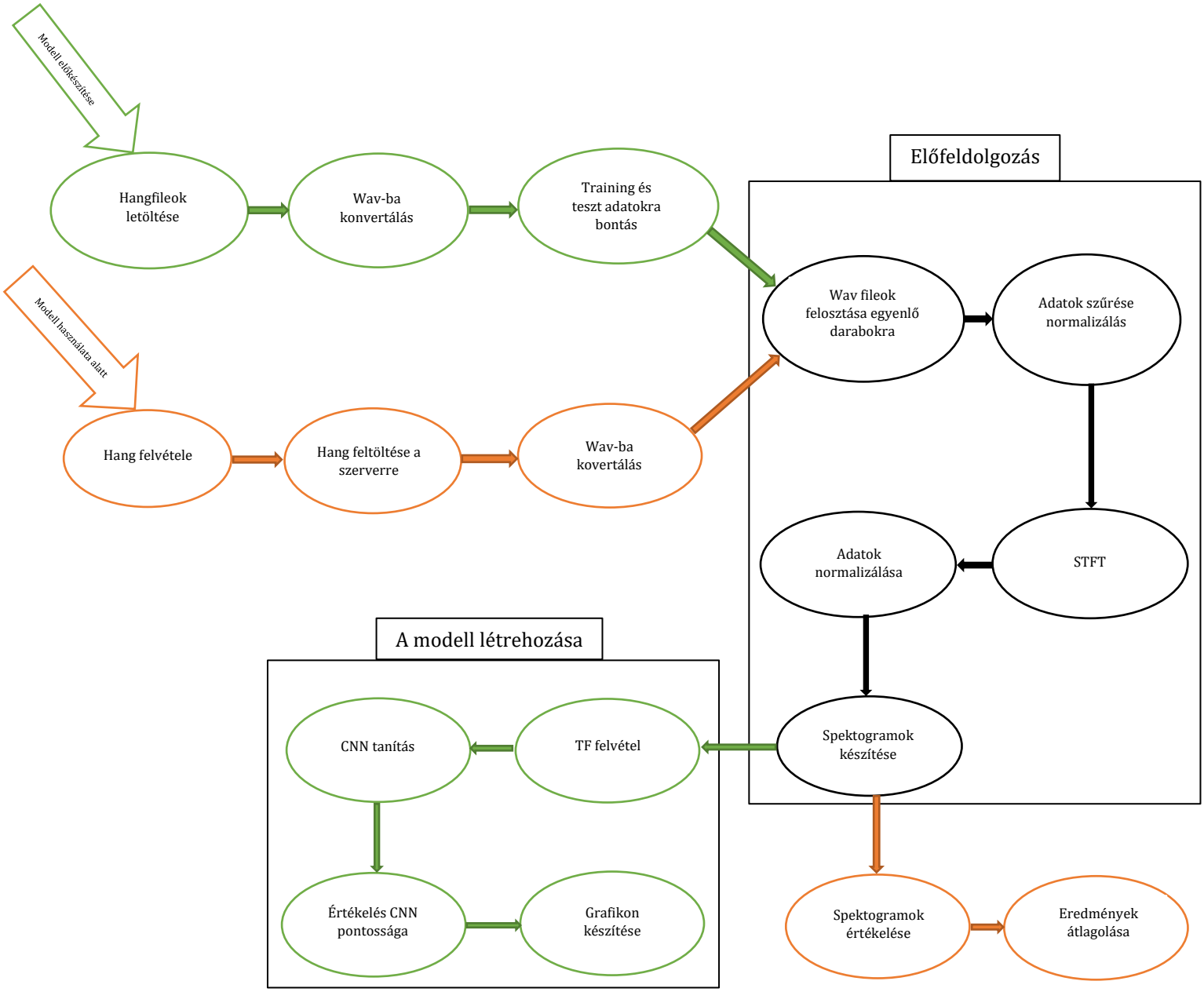
2. ábra Mel skála

ahol m a normál frekvencia Mel-skálája.

A hangfájlok(Wav) egyenlő hosszúságuak és normalizáltak.



3. ábra STFT és MFCC képek létrehozásának szakaszai.



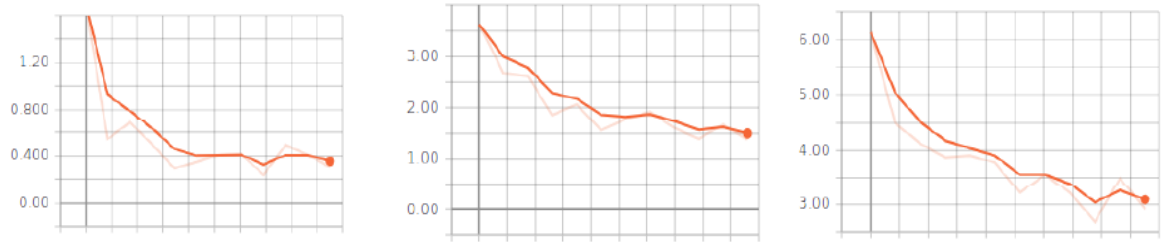
4. ábra Model folyamatábrája előkészítés (zöld),értékelési szakasz (narancs), előkészítés előfeldolgozás (fekete).

Konklúzió

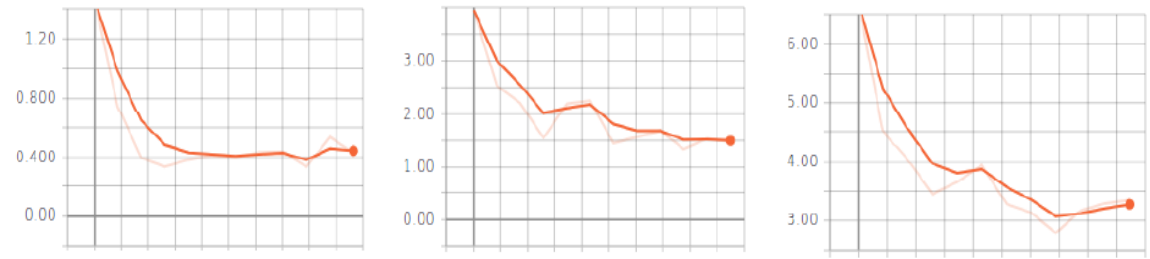
A madárhangfelismerés a gépi tanulás segítségével folyamatos növekedést mutatott az elmúlt években, a kezdetektől fogva rengeteg időt fordítottak különböző neurális hálók betanítására [8].

Alternatív és esetleg másik lehetséges megoldás a, transzfertanulás. Finomhangolás, előre betanított hálózat, a hang vizuális megjelenítésével, spektrogramok bemutatásával. MobileNet hálózat képei és adatai sokban különböznek az alacsonyabb szintű rosszabb minőségű képektől. A kezdeti kísérletek azt mutatják, hogy eredményes pontosság csak 2 osztály használatával érhető el [9].

A kísérletek összehasonlítják a színes képekkel készült spektogramokat is. Az eredmények arra engednek következtetni, hogy az RGB spektogramok sokkal hatékonyabbak mint a fekete fehér társai. A különbség a színes és a szürkeárnyalatos pontossága között folyamatosan növekszik, mikor több osztályt használunk [8].



5. ábra Balról jobbra 2 osztályos Jet színtérkép, 10 osztályos Jet, és az utolsó az 50 osztályos Jet



6. ábra Balról jobbra 2 osztályos szürkeárnyalatos, középen 10 osztályos szürkeárnyalatos a jobb szélén pedig az 50 osztályos szürkeárnyalatos

Az 5-ös, 6-os ábra a tanítás során fellépő veszteségeket mutatja be.

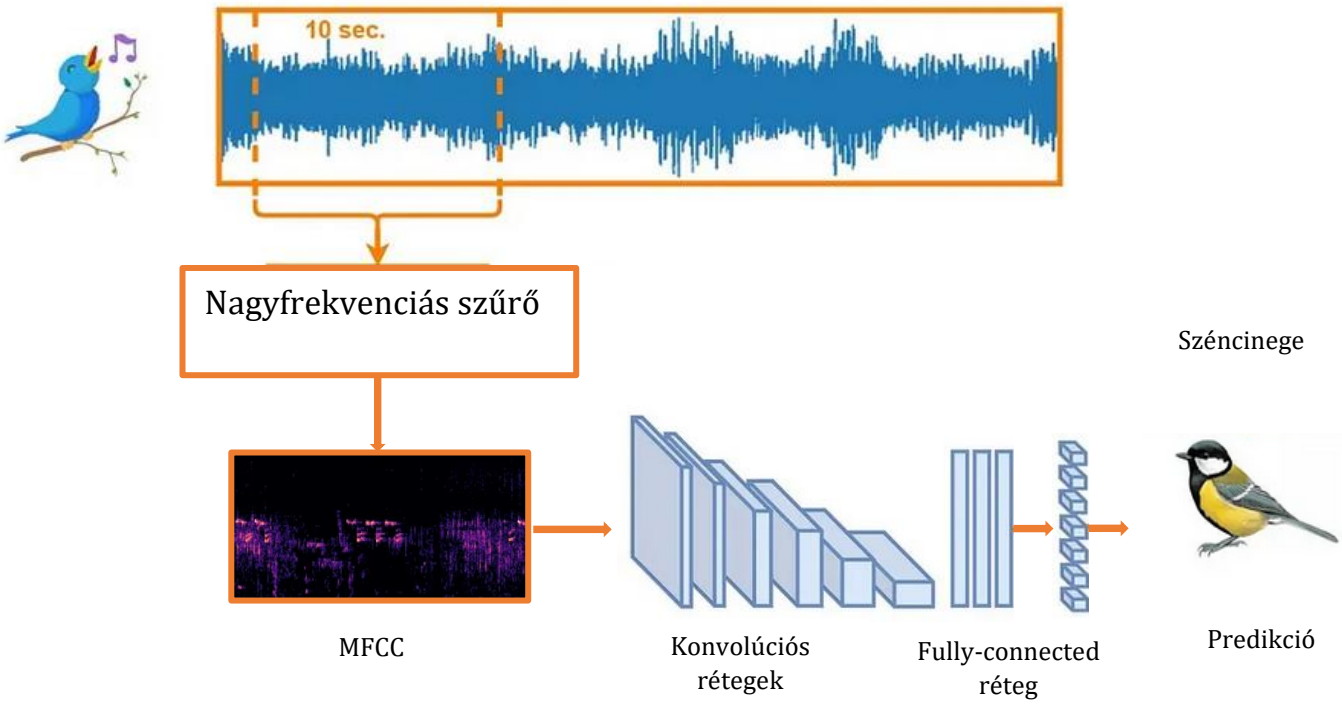
A nagyobb pontosság elérése érdekében a következő fejlesztéseket érdemes megfontolni. Egy nagyobb, robusztusabb előre betanított konvolúciós neurális hálózat, mint például a ResNet segít a pontosság és a nagyobb számok elérésében.

Megfelelően megkülönböztetni és összehasonlítani az RGB és a szürkeárnyalatos spektrogramokat, a fekete fehér spektrogramok esetében az alsó rétegeket kell a szürkeárnyalatos képekre tanítani [9].

Továbbiakban az előfeldolgozási szakaszban javasolt a zajcsökkentés, valamint szűrés az extrém frekvenciák esetében. Különböző gamma értékek tesztelése növelheti a hasznos információk mennyiségét a spektrogramok esetében [8].

A pontosság a modell minőségi mérése, ami azt mutatja meg, hogyan lenne képes a modell újra lefutni, új adatokkal, melyeket korábban nem használt [10].

$$\text{Pontosság} = \frac{\text{Helyes predikciók száma}}{\text{Teljes predikciók száma}}$$



7. ábra hangadatok előfeldolgozása és neurális hálózat modellje

Hivatkozások

[1] Priyadarshani, N., Marsland, S., Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: A review. Journal of Avian Biology, 49(5): jav-01447.

[2] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol. 24, pp. 279–283, 2017.

[3] T. Giannakopoulos and A. Pikrakis, Introduction to Audio Analysis: a Matlab approach. 2014.

[4] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," IEEE J. Sel. Top. Signal Process., vol. 13, no. 2, pp. 206–219, 2019.

[5] M.A. Imtiaz and G. Raja, "Isolated Word Automatic Speech Recognition (ASR) System Using MFCC, DTW & KNN," Proc. - APMediaCast 2016, pp. 106–110, 2017.

[6] Baba, T. (2012). Time-frequency analysis using short time Fourier transform. The Open Acoustics Journal, 5(1): 32-38.

[7] Lee, C.H., Han, C.C., Chuang, C.C. (2008). Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. IEEE Transactions on Audio, Speech, and Language Processing, 16(8): 1541-1550.

[8] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks," Working notes of CLEF, 2017.

[9] K. J. Piczak, "Recognizing bird species in audio recordings using deep convolutional neural networks," Working notes of CLEF, 2016.

[10] Baratloo, A., Hosseini, M., Negida, A., El Ashal, G. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Emergency (Tehran, Iran), 3(2): 48-49