PROYEK DEEP LEARNING
"DETEKSI ALAT MUSIK DALAM REKAMAN MENGGUNAKAN MEL-SPECTROGRAM, MFCC DAN CNN"

OLEH:

Alvin Christian Lorence	**(C14180045)**
Geraldino Dharmawan	**(C14180018)**
Verick Gozali	**(C14180060)**
Sheeren Anggela	**(C14180070)**
Andrianto Saputra	**(C14180104)**
Vincent Kurniawan	**(C14180191)**

UNIVERSITAS KRISTEN PETRA

TEKNIK INFORMATIKA

2021

1. Introduction

Suara merupakan getaran yang merambat melalui gelombang akustik dan bisa dihitung dengan satuan frekuensi. Suara ada di sekitar lingkungan yang berisi sejumlah informasi pada peristiwa yang terjadi di dekatnya, pada umumnya manusia mampu mengenali dan mendengar banyak peristiwa suara.

Mesin pengolahan suara otomatis yang yang canggih masih belum sempurna untuk mengenali suara secara sempurna, hal ini membutuhkan penelitian lebih lanjut yang diperlukan untuk mengembangkan sistem yang canggih dan tahan lama yang mampu mengenali berbagai peristiwa dalam getaran suara pada aliran audio yang realistis.

Secara khusus, tujuan dari tugas ini adalah untuk membangun sistem penandaan audio yang dapat mengkategorikan klip audio sebagai milik salah satu dari kumpulan 41 kategori beragam yang diambil dari AudioSet Ontology (terkait alat musik, suara manusia, suara domestik, binatang,dll). Salah satu motivasi untuk tugas ini datang dari yang besar jumlah konten audio buatan pengguna yang tersedia di web, yang dapat menjadi sumber daya yang sangat potensial untuk pengenalan suara terkait penelitian.

Penggunaan data tersebut untuk melatih sistem penandaan audio sehingga menimbulkan masalah. Secara khusus, tugas ini berkaitan dengan audio yang dibuat pengguna pada klip yang diambil dari Freesound, yang sangat beragam dalam hal konten akustik, teknik perekaman, durasi klip, dll. Demikian juga,klip audio ini terkadang menampilkan tidak lengkap dan tidak konsisten pada metadata yang disediakan pengguna. Untuk menyiapkan dataset untuk tugas ini, beberapa klip audio diberi label secara manual menggunakan subset dari 41 kategori, sementara kumpulan klip yang lebih besar secara otomatis dikategorikan pada dasar metadata yang disediakan pengguna yang ada.

2. Dataset

Dataset yang kami gunakan adalah dataset Freesound Dataset Kaggle 2018 (FSDKaggle2018) yang digunakan pada kompetisi Freesound General-Purpose Audio Tagging Challenge. Dataset ini terdiri dari 11.703 buah file audio yang dikategorikan menjadi 41 label dengan total ukuran file sebesar 10 GB. Tipe rekaman yang digunakan adalah dengan format mono uncompressed PCM 16-bit .wav dengan audio rate sebesar 44.1 kHz yang diambil dari website Freesound.org.

Data dalam dataset ini dibagi menjadi 2 bagian, yaitu train set dan test set. Pada train set, terdapat kurang lebih sebanyak 9.500 sample audio yang memiliki panjang dan ukuran audio yang beragam, dari 300ms hingga 30 detik. Total panjang audio pada train set adalah 18 jam. Sedangkan pada test set, data yang tersedia ada sebanyak 1.600 sample dengan total panjang audio 2 jam.

Semua file audio pada data set ini masing-masing hanya diidentifikasi dengan 1 label, dan proses anotasi (labeling) pada audio dilakukan secara otomatis dengan verifikasi manual. Data label yang digunakan diambil dari AudioSet Ontology, pada proses verifikasi kurang lebih sebanyak 3.710 (31.7%) sample dari data train dan 1600 (100%) dari data test telah diverifikasi secara manual dan dikategorikan dengan data label tersebut. Untuk sisa dari data yang belum terverifikasi, diperkirakan sebanyak 60 hingga 70 persen dari sisa data tersebut memiliki anotasi yang benar atau sesuai dengan audio file. (Fonseca et al., 2019)

3. Metode

Metode yang digunakan adalah CNN (Convolutional Neural Network). CNN (Convolutional Neural Network) adalah metode deep learning yang merupakan salah satu class dari deep neural network biasanya digunakan untuk menganalisis visual imagery alias gambar.

Cara kerja CNN sendiri adalah mereka menggunakan multi-channeled images. Tidak seperti flat image yang manusia lihat yang hanya bisa melihat lebar dan tinggi

CNN tidak bisa mengenali itu. CNN menggunakan 3 warna yaitu Red-Green-Blue encoding, CNN mencampur 3 warna tersebut untuk menghasilkan warna yang manusia tangkap

Convolutional Network menyerap gambar seperti tiga lapisan warna terpisah yang ditumpuk satu diatas yang lainnya. Gambar berwarna normal terlihat sebagai kotak persegi panjang yang lebar dan tingginya diukur dengan jumlah piksel dari dimensi tersebut. Lapisan kedalaman dalam tiga lapisan warna(RGB) yang ditafsirkan oleh CNN disebut sebagai *Channel*(saluran).

Layer pertama dari CNN itu disebut dengan Convolutional layer, dimana layer ini adalah inti dari CNN itu sendiri dan yang paling bekerja keras. Data atau gambar yang ada digulung menjadi satu menggunakan filter atau kernel. Filter adalah unit yang kecil yang di apply di dalam data melewati sliding window.

Layer kedua dari CNN disebut dengan Activation Layer yang menerapkan ReLu(Rectified Linear Unit), di layer ini juga diterapkan fungsi rectifier untuk meningkatkan non-linearitas di CNN. Gambar dibuat dari objek yang berbeda, dimana tidak linear satu sama lain.

Layer ketiga dari CNN disebut dengan Pooling Layer yang melibatkan fitur downsampling. Biasanya diterapkan melalui setiap lapisan dalam volume 3d. Terdapat juga hyperparameter di dalam lapisan ini :

1. Dimensi perluasan spasial : Merupakan nilai n yang dapat diambil N crossnya dan representasi fitur serta petakan ke nilai tunggal.
2. Stride : Merupakan berapa banyaknya fitur yang dilewati sliding window sepanjang lebar dan tinggi.

Layer yang keempat dan terakhir adalah Fully Connected Layer yang melibatkan Flattening(Perataan). Selain Flattening,layer ini juga melibatkan transformasi seluruh matriks peta fitur yang dikumpulkan menjadi satu kolom yang kemudian diumpankan ke jaringan saraf untuk diproses. Dengan adanya layer keempat ini, fitur-fitur di layer ini dapat digabungkan bersama-sama untuk membuat model. Setelah proses [penggabungan tersebut, dapat dimiliki fungsi aktivasi seperti softmax atau sigmoid untuk mengklasifikasikan output. (Bansari, 2019)

Kelebihan dari menggunakan algoritma CNN (Convolutional Neural Network adalah CNN akan secara otomatis mengetahui fitur mana yang penting tanpa supervisi seorang manusia. (Dertat, 2017)

Sedangkan kekurangan yang kita dapatkan dari menggunakan CNN adalah algoritma ini sedikit lebih pelan dikarenakan max pool nya. Jika CNN memiliki beberapa layer maka training process nya akan mengambil banyak waktu jika tidak mempunyai GPU yang bagus. CNN membutuhkan data set yang besar untuk memproses dan training neural network tersebut (Bhuiya, 2021)

4. Pengujian

Rancangan pengujian yang akan dilakukan untuk deteksi alat musik dalam rekaman menggunakan metode CNN dan fitur MEL-Spectrogram dan MFCC. Metode ini digunakan untuk dapat melihat perbandingan akurasi yang ada dalam dataset apakah ada perbandingan yang banyak atau tidak. Untuk fitur MFCC sendiri ini merupakan salah satu metode yang banyak digunakan dalam bidang *speech technology*, baik *speaker recognition* maupun *speech recognition*. Fitur ini digunakan untuk melakukan *feature extraction*, sebuah proses yang mengkonversikan sinyal suara menjadi beberapa parameter. Sedangkan pada spectrogram, sinyal suara dikonversikan dalam sebuah value array yang kemudian dapat divisualisasikan dalam bentuk graph atau chart.

Untuk membuat model, kami menggunakan 2 jenis layer yang berbeda sebagai patokan perbedaan hasil. Perbedaan dari kedua jenis layer yang kami gunakan terletak pada jumlah layer MaxPooling2D, Dense dan Dropout, namun kedua layer tersebut memiliki weight yang sama. Tabel perbedaan dari kedua jenis layer adalah sebagai berikut:

Layer Jenis 1	Layer Jenis 2
Conv2D (16,(3,3),'relu')	Conv2D(16,(3,3),'relu') MaxPooling2D(2,4)

Conv2D(32,(3,3),'relu') MaxPooling2D(2,2) Dropout(0.5) Flatten() Dense(128) Dropout(0.5) Dense(64,'relu') Dropout(0.5) Dense(5)	Conv2D(16,(3,3),'relu') MaxPooling2D(2,2) Flatten Dense(64,'relu') Dropout(0.25) Dense(5)

Masing-masing dua jenis layer tersebut kami train dengan 5 class dari keseluruhan 41 yang tersedia dalam dataset yang kami gunakan. Alasan kami meminimalisir jumlah data yang digunakan adalah karena pada saat pengujian, hasil training dari kedua jenis model dengan 41 class memiliki akurasi yang rendah, yaitu sekitar 35 hingga 40%. Ketika kami mencoba melakukan training ulang, resource yang digunakan terlalu besar sehingga membuat komputer training utama kami mengalami kegagalan sistem fatal sehingga crash dan kemudian mendapatkan error BSOD. Berdasarkan kejadian tersebut, akhirnya kami memutuskan untuk men-training model dengan 5 data untuk menghemat resource dan waktu serta meningkatkan akurasi.

Untuk proses predicting dengan menggunakan raw data, ada limitasi terhadap data yang bisa digunakan untuk meningkatkan efektivitas proses prediksi menggunakan model yang kami train. Dataset yang kami gunakan untuk men-training model memiliki audio dengan panjang bit sebesar 16-bit, sehingga untuk raw data yang digunakan untuk proses prediksi harus memiliki minimum panjang bit yang sama.

5. Hasil

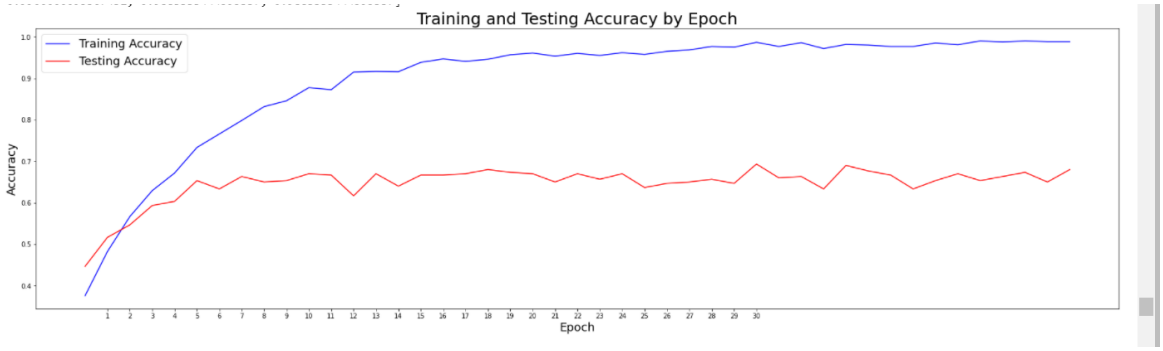
Tingkat dan perbandingan akurasi Dengan Dataset 5 Macam instrumen

Epoch : 30	MFCC	SPECTROGRAM

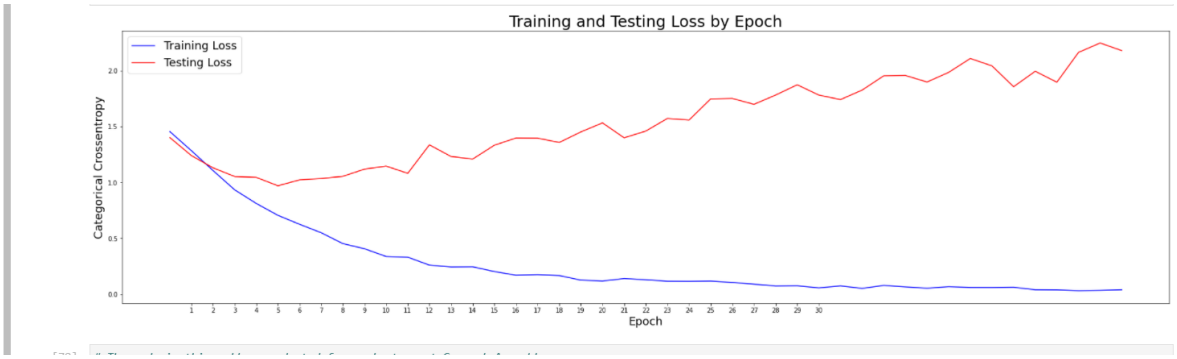
Model Layer 1	83%	95%
Model Layer 2	74%	98%

Jadi dari hasil yang kami temukan bahwa dengan model layer 1 fitur MFCC dapat mendapatkan model layer 1 yang mempunyai lebih banyak layer dan hanya menggunakan dua layer convutional 2D dibandingkan dengan jenis Layer ke-2. Pada Fitur spectrogram terjadi overfitting pada data trainnya dari 95% menjadi 68% pada layer 1 dan pada layer 2 dari 98% menjadi 65% serta val loss yang jauh membuat kami menggunakan fitur mfcc sebagai solusi yang terbaik dibandingkan untuk layer 1 fitur mfcc mempunyai val_acc 85%. Dalam layer 2 fitur mfcc didapatkan 76% accuracy dan val_accuracynya 76%

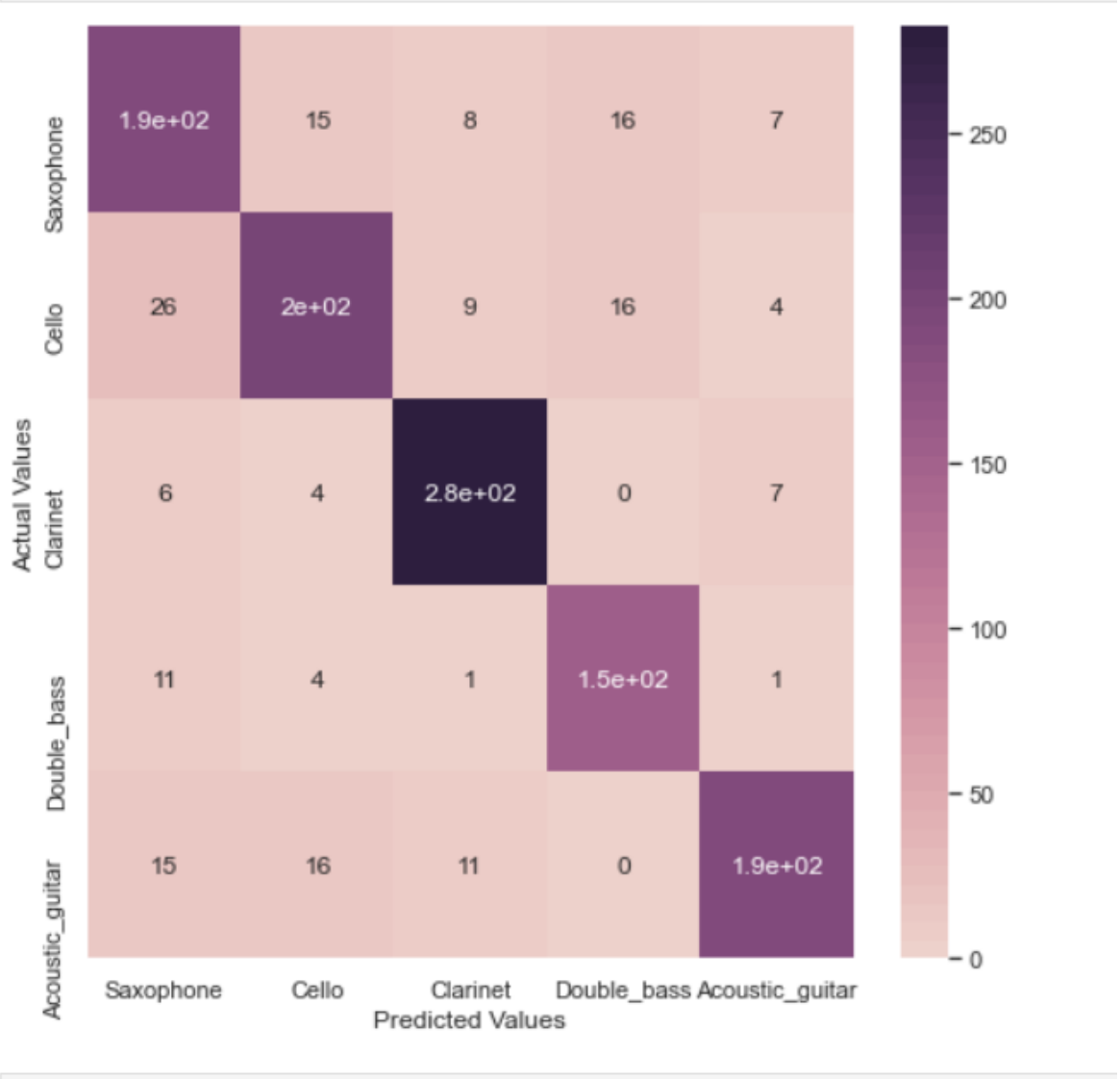
Value Loss Layer dari layer 1



Value Loss Layer dari layer 2

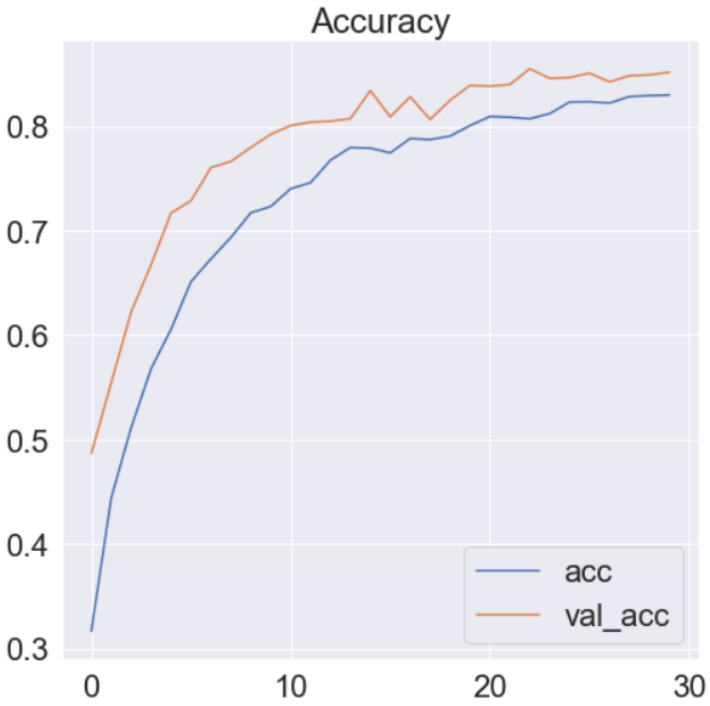


Maka kami menggunakan model layer 1 dengan fitur MFCC untuk mendeteksi jenis instrumen yang akan di testing dengan nilai default dari MFCC, yaitu:40 serta sample rate: 6000 dan mendapatkan akurasi 83% dengan confusion matriks yang sudah ditrain sebagai berikut:

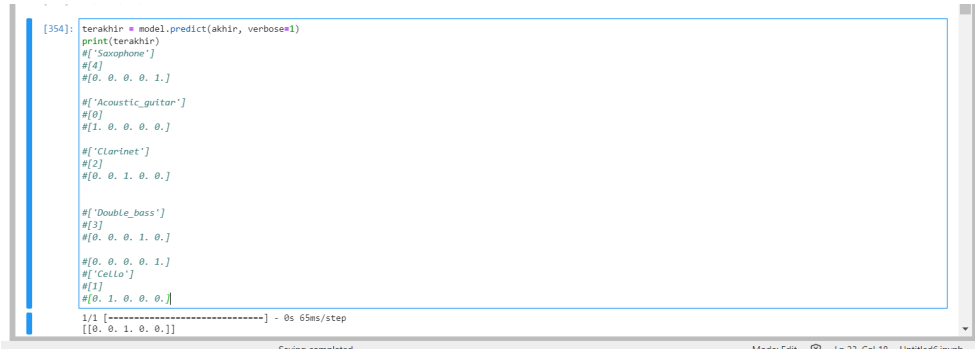


Serta dilihat dari grafik testing dan data validasi yang didapatkan bahwa tidak terjadi underfitting maupun overfitting dari model yang sudah dibuat maka dengan tingkat akurasi 82% dan 85% untuk validasi data

```
acc: 0.8299832344055176
val_acc: 0.8517587780952454
```



Untuk contoh hasil prediksi kita memakai model ini dengan data yang sudah kita coba sesuai dengan cara yang sama :



```
[354]: terakhir = model.predict(akhir, verbose=1)
print(terakhir)
#['Saxophone']
#[4]
#[0. 0. 0. 0. 1.]

#['Acoustic_guitar']
#[0]
#[1. 0. 0. 0. 0.]

#['Clarinet']
#[2]
#[0. 0. 1. 0. 0.]


#['Double_bass']
#[3]
#[0. 0. 0. 1. 0.]

#[0. 0. 0. 0. 1.]
#['Cello']
#[1]
#[0. 1. 0. 0. 0.]

1/1 [==============================] - 0s 65ms/step
[[0. 0. 1. 0. 0.]]
```

6. Kesimpulan

Data yang kami gunakan dalam tugas ini diambil dari FSDKaggle2018 yang menyajikan kekhasan memiliki subset data pelatihan dengan anotasi berbagai keandalan serta menampilkan klip audio berdurasi variabel yang hanya menggunakan 5 label musik, yaitu saxophone, clarinet,hi bass, cello, dan acoustic guitar. Rancangan pengujian untuk deteksi alat musik dalam rekaman menggunakan metode CNN dan fitur MEL-Spectrogram dan MFCC. Pembuatan

model, kami menggunakan 2 jenis layer yang berbeda sebagai patokan perbedaan hasil. Perbedaan dari kedua jenis layer yang kami gunakan terletak pada jumlah layer MaxPooling2D, Dense dan Dropout, namun kedua layer tersebut memiliki weight yang sama. Dengan data yang kita training sebanyak 30 epoch untuk mencoba memakai spectogram dan MFCC kami mendapatkan hasil terbaik dari dua macam model layer dengan tidak overfit maupun underfit dimana mencapai 85%. Pada fitur dan model layer yang lain kami melihat dengan layer 1 dengan feature mfcc adalah yang terbaik

7. Daftar Pustaka

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, Manoj Plakal, Daniel P. W. Ellis, & Xavier Serra. (2019). FSDKaggle2018 (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2552860

Fonseca, E., Plaka, M., Font, F., Ellis, D. P. W, Favory, X., Pons, J., Serra, X.(2018). General-Purpose Tagging of Freesound Tagging of Freesound Audio with Audioset Label : Task Description, Dataset, and Baseline. Barcelona Journal of Music Technology Group, 1-6. Retrieved from https://arxiv.org/pdf/1807.09902.pdf

Bansari, S., (February 13, 2019). "Introduction to how CNNs Work". https://medium.datadriveninvestor.com/introduction-to-how-cnns-work-77e0e4cde99b