大数据与机器智第四次作业

Homework 4 , Mel Spectrogram Classification

姓 名: 孟 念 Muennighoff Niklas
学 号: 1800092850
班 号: 01510243-90

大数据与机器智能
(秋季, 2021)

清华大学
Zhen Chen、Min Guo、Yisong Zhang
Hang-gao Xin、Ya-ning Guo、Zi-yi Zhang

2021 年 12 月 5 日

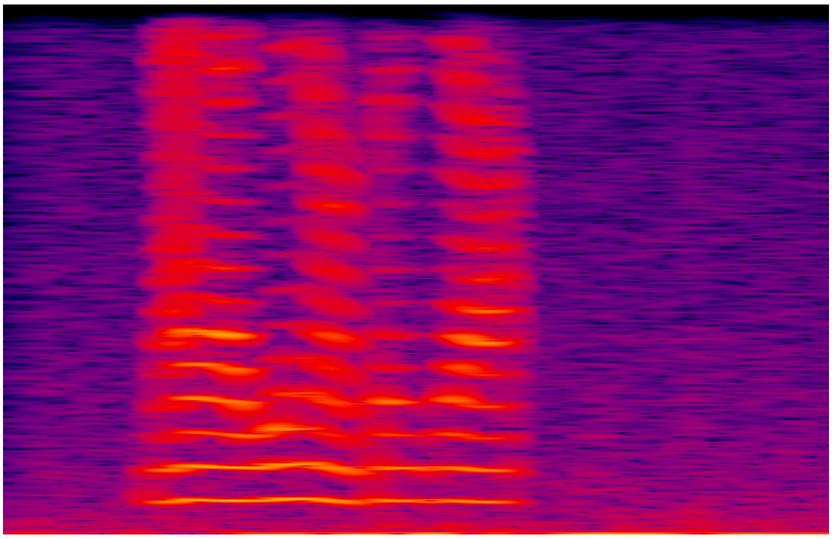
目 录

1 Introduction

This work constitutes the fourth homework of the BDMI class. The task is to classify mel spectrograms. The report consists of the following parts:

- **Data**: In this module, I will briefly go over the data type, its distribution & augmentations.
- **Modelling**: In this module, the used models, hyperparameter settings & results will be explained.

Below is an example mel spectrogram from the provided dataset.

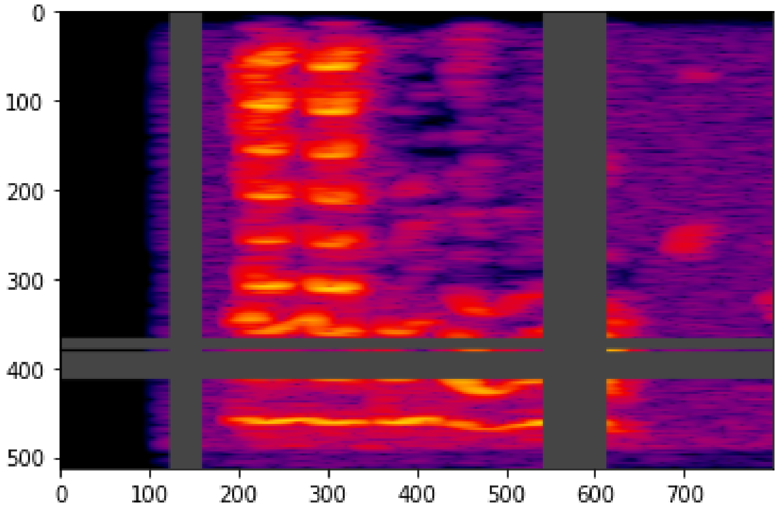


2 Data

The data consists of mel spectrogram images that have previously been generated from human voices. The data has a total of 2469 samples. They're grouped into 24 classes. This means there are only about 100 samples per class.

At this stage, I realized that training a performant 24-class classification neural network on just 100 samples will be considerably difficult. Hence, I decided to use as much data as possible for training. I splitted the data into a 90% train and a 10% test set. I skipped a separate validation set due to the few data. I havn't performed extensive optimization based on test results. This makes for a total of 2222 train and 247 val / test images.

I use two augmentation techniques: Frequency masking & time masking. Find below an example of an augmented mel spectrogram.



Frequency corresponds to the y-axis of the image. Frequency-masking consists of randomly inserting horizontal lines to hide different frequencies from the model. Time corresponds to the x-axis of the image. Time masking are hence the vertical lines masking out random time segments.

3 Models

I have used two different CNN models to classify the images. The first is a pretrained ResNet, while the second is a custom CNN trained from scratch.

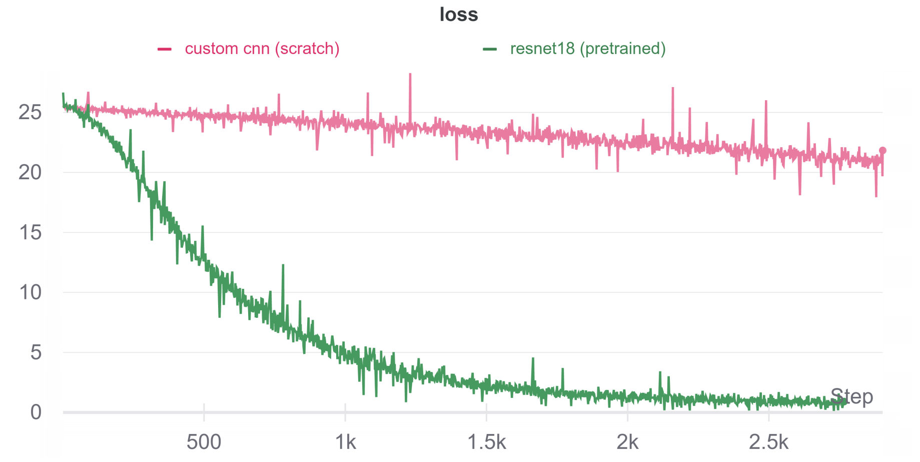


图 1: Training loss - Blue corresponds to the pretrained resnet, orange to the CNN from scratch

3.1 ResNet (pretrained)

I use a pretrained resnet18 [1], replace the final linear layer and finetune all parameters. I did not perform extensive hyperparameter tuning and just train for 500 epochs on the entire training set with an Adam optimizer and a learning rate of 1e-4. The model converges after around 20,000 steps or roughly 100 epochs.

3.2 Custom CNN (scratch)

I also train a simple CNN from scratch consisting of four convolutional layers with ReLU activation and BatchNorm. The model is trained using the same optimization scheme as the pre-trained model.

3.3 Comparison

Figure 1 shows the training and 2 the validation loss curve. Overall, both model variants converge, as their loss trends down. However, the pretrained variant trains much faster reaching a minimum after around 2k steps (corresponding to 40k gradient updates).

Figure 3 shows the training and validation accuracy. Figure 4 shows the training and validation F1 scores. F1 is less influenced by an unbalanced dataset, hence a good metric to check in tandem with accuracy for multi-

[1] Deep residual learning for image recognition, He et al.

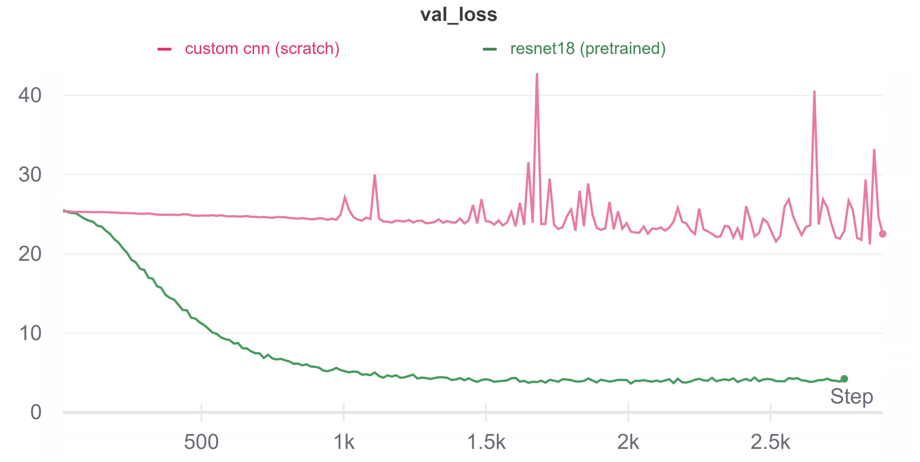


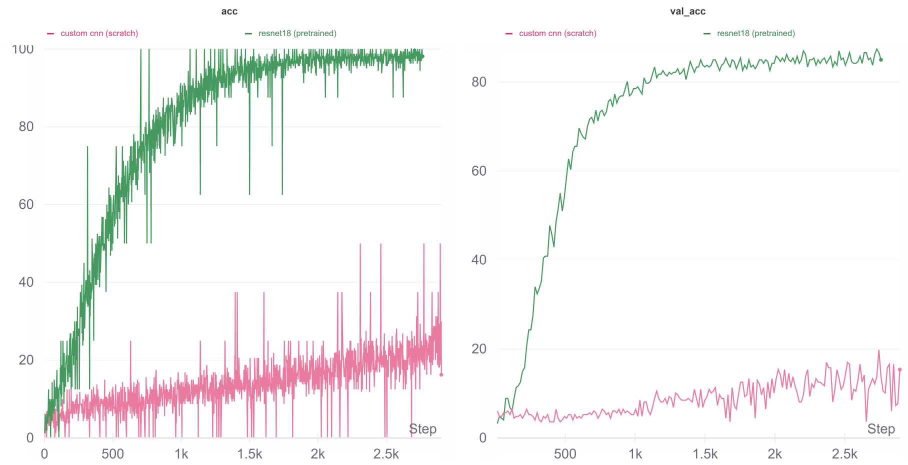
图 2: Validation loss - Blue corresponds to the pretrained resnet, orange to the CNN from scratch

class problems. The graphs show similar trends as the loss curves. The pretrained resnet reaches a maximum validation accuracy of around 85%. The training accuracy is around 98% towards the end. The model hence slightly overfit to the training set. Using more regularization could balance the performance on the train and val sets.

4 Learnings

In this report, I have presented my mel spectrogram classification results. Throughout this project I have learnt:

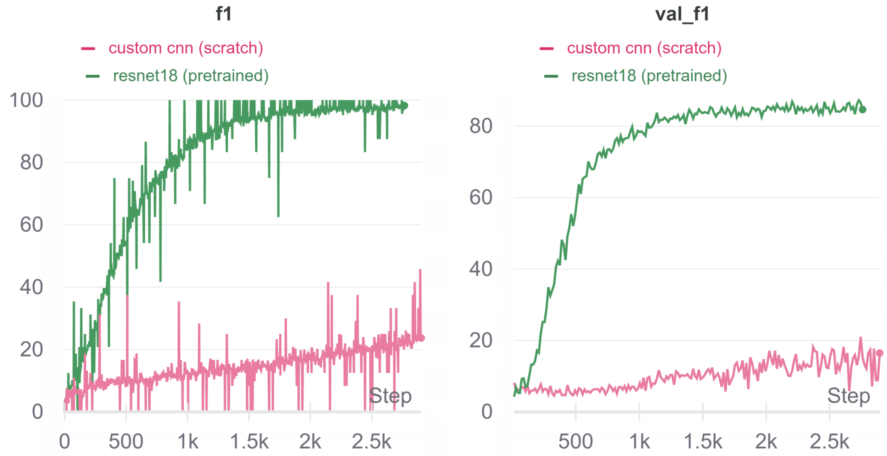
- What are Mel spectrograms
- How to perform augmentation on Audio
- Sometimes just let the model train for a while & eventually it will converge :)



(a) Training accuracy

(b) Validation accuracy

图 3: Accuracy - Blue corresponds to the pretrained resnet, orange to the CNN from scratch



(a) Training f1

(b) Validation f1

图 4: F1 - Blue corresponds to the pretrained resnet, orange to the CNN from scratch