Melody Extraction

Aman Verma
180075
Department of Electrical Engineering
amanve@iitk.ac.in

Deshpande Prasad Jayant
17103270
Department of Civil Engineering
prasadj@iitk.ac.in

Raj Prakash Gohil
19104407
Department of Electrical Engineering
rgohil@iitk.ac.in

Abstract

This paper discusses a CREPE: A Convolutional Representation for Pitch Estimation algorithm as a baseline model used for melody extraction. The model algorithm is trained on MIR-1K and tested on MIR-50K dataset. The highlights and limitations of CREPE are discussed. To overcome the limitations, another algorithm will be tested in future which is Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness.

1 Introduction

Music Information Retrieval (MIR) is an interesting yet challenging task in audio signal processing. It includes pitch extraction, beat tracking, tempo estimation, source separation etc. The end-user applications of MIR include searching-by-humming, searching-by-example, searching-by-notes etc. It may also include searching songs based on mood or genre. [1]

CREPE algorithm is a data-driven algorithm developed in 2018, which is highly precise, maintaining over 90 percent raw pitch accuracy (RPA) even for a strict evaluation threshold of just 10 cents. [2]. Its Python implementation along with a pre-trained model makes it easy to utilize and reproducible. In this report, the ready made library and pre-trained model is not used. The further details about CREPE are mentioned in Methodology section.

The CREPE algorithm will act as a reference i.e. baseline model for further algorithm development.

2 Problem definition

Melody extraction problem can be defined in 2 ways -

a) Identification of dominant frequency of the music (which in measured in Hz). This problem is mathematically modelled as a regression problem.

b) Identification of Swaras/Notes of the music as a Classification problem. In this case, the western notes in terms of C, G etc. are found out in terms of absolute frequency. Indian music Swaras in terms of Sa, Re, Ga etc. are found out w.r.t. a suitable reference frequency.

In many music forms (e.g. Indian classical music), the notes played do not belong to exact location of the notes, but there are little perturbations (Vibrato / Murki), or continuous shift of one note to another note (slide). In such cases, deterministic identification of notes is not the best way to

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

identify notes, but one can identify them probabilistic way (e.g., in case of CREPE). In this case, uncertainty or confidence interval associated with the prediction. Wherever the uncertainty is high, human inspection can be carried out to identify the melody. Hence, uncertainty analysis will become an input to active learning.

In the present study, we have used both approaches for melody extraction along with uncertainty analysis.

MIR1K dataset is used for training-testing-validation.

3 Methodology

3.1 Gaussian Processess

Formally, a Gaussian Process is defined as a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions. A Gaussian process therefore defines a distribution over functions and is fully specified by a mean function $m(x)$ and a covariance function $k(x, x')$. Write $f \sim GP(m, k)$ meaning f is distributed as a GP with mean *m* and covariance k. Given a finite number n of locations $\boldsymbol{x} = [x_1, ..., x_n]$ it is possible to sample from the GP at those locations. First compute $\mu_i = m(x_i)$, $\sum_{i,j} = k(x_i, x_j)$ and then sample a vector from the distribution $\boldsymbol{f} \sim \boldsymbol{N}(\boldsymbol{\mu}, \sum)$

3.2 Regression approach

Amersfoort et. al. in their recent (Feb 2021) paper 'Improving Deterministic Uncertainty Estimation in Deep Learning for Classification and Regression' [3] proposed a method that provides uncertainty estimates. This model that estimates uncertainty in a single forward pass and works on both classification and regression problems. For uncertainty estimation, a wide ResNet is used as a feature extractor which becomes input to Gaussian Process(GP). Variational approximation to the full GP is made. Cost function used is ELBO. Along with usual hyperparameters of neural networks, there is a step of spectral normalization. We have tried different kernals and number of inducing points for variational GP.

For regression, the output is given as prediction (mean of the GP outputs) and uncertainty (variance of the GP outputs).

We have tried the following inputs strategies as below -

a) Using time-domain input The audio input is sampled at 16kHz frequency and a set of 1024 samples in given as input to the input layer.

b)Using STFT input STFT is carried out on the input audio signal, keeping NFFT as 63 and hop size as 63 so that the input gets 32 features.

Following two models we have tried as a feature extractor, both were provided by the author and we have modified them

a) FCResNet A lightweight fully connected ResNet

b)WideResNet A comparitively heavier convolutional ResNet which was originally provided for classification CIFAR10 dataset. We have modified it for regression problem. Here consecutive samples are appended together to make a Nx32x32 shape input, where N is number of samples.

3.2.1 Accuracy evaluation

For visual monitoring line plots, scatter plots of ground truth frequencies and predicted values are observed. Moreover, R^2 value, coefficient of correlation, RMSE was calculated.

3.3 Classification approach

Predicted pitch is said to be correct (or within acceptable limits) if it lies within half semitone above or below the ground truth semitone. This criteria is used to calculate the RPA (Raw Pitch Accuracy).

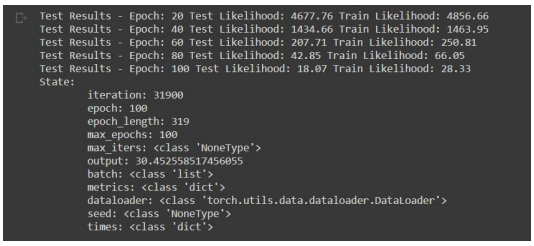

Figure 1: Training performance

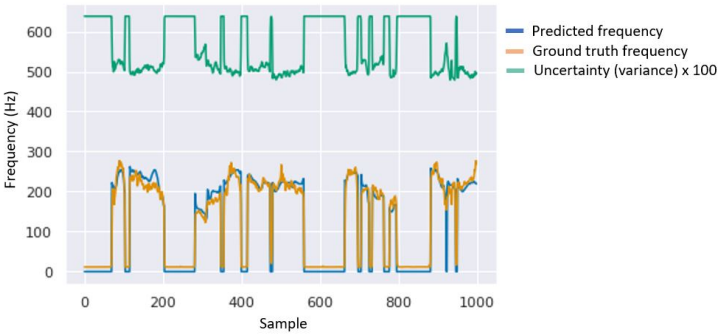

Figure 2: Predicted and GT frequencies along with uncertainty (x100)

In classification tasks, the last layer is often passed through a softmax activation which returns a probability distribution over the classes.

However, a high probability doesn't mean that the model is certain of its prediction: the model could give a high probability with a high uncertainty in which case the prediction might be worthless. So to overcome this, we experimented with two different methods:

- Obtain uncertainty bounds for convolutional neural networks: using a Gaussian process on the last-layer features learned by the CNNs. Using a Gaussian process trained on the features extracted by a CNN that was trained in a previous step on the same test training data.
- Using Bayesian convolutional neural networks with variational inference, a variant of convolutional neural networks (CNNs), in which the intractable posterior probability distributions over weights are inferred by Bayes by backpropagation. For this we implemented BBB or Bayes by Backpropagation type of layer, i.e., this layer samples all the weights individually and then combines them with the inputs to compute a sample from the activations.

4 Results and discussion

4.1 Results for regression

4.1.1 WideResNet

The results for WideResNet without spectral normalization are shown in Figures 1, 2, and 5. Figure 1 shows that WideResNet is able to model the training data well (Coeff. of correlation is 0.98). Figure 2 shows corresponding training performance. Figure 3 suggests that model is predicting non-zero frequencies where the GT corresponds to 0, but from Figure 1 we can see that the uncertainty value is maximum at such locations. The model was given almost constant output with spectral normalization, we could not find the reason.

4.1.2 FCResNet

Though we tried different model configurations and hyperparameters but the model was not able to predict the frequencies correctly. As seen in Figure 4, the model is not able to predict the values above 200 Hz. This bound increases along with number of epochs, but on the other hand performance on validation data decreases i.e. overfitting.

4.2 Results for classification

We got around 42% accuracy for the Gaussian process method when using Matern32 Kernel along with White noise and for making training possible, we used GPFlow, which implements various approximation algorithms for Gaussian Processes. We used the Sparse Variational Gaussian Process Classifier. The model learns a set of m inducing points which are used instead of the training points in

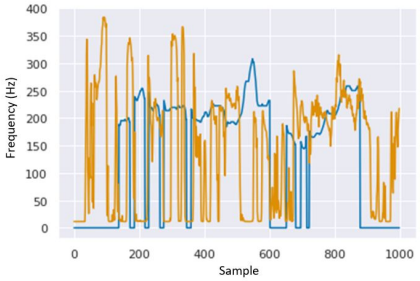

Figure 3: Predicted and GT frequencies for WideResNet

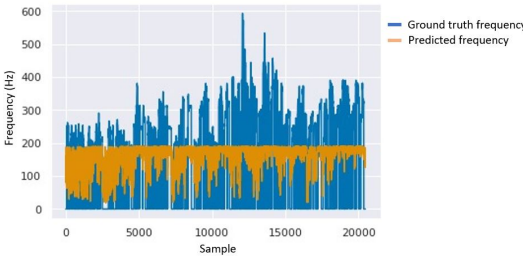

Figure 4: FCResNet predicted and GT frequencies

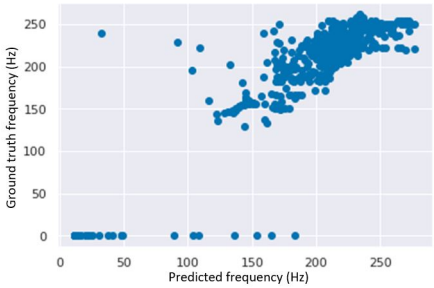

Figure 5: Scatterplot of predicted and GT values

the Gaussian process. Typically, m < n, which makes the task tractable as its complexity is $O(nm^2)$. But when using the Bayesian approach, we get nearly 67.42% validation accuracy when trained on 10 epochs.

5 Conclusion

The GP based deep learning methods have potential to estimation the prediction uncertainty [4, 3]. This estimation can be further used for various other AI applications,e.g. active learning. Moreover, the proposed models need to be improved in terms of accuracy. A deep understanding of the dataset characteristics and GPs may be helpful for tuning such models.

References

[1] Roberto Raieli. 6 - the current status of mir systems. In Roberto Raieli, editor, *Multimedia Information Retrieval*, Chandos Information Professional Series, pages 175–193. Chandos Publishing, 2013.

[2] J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018.

[3] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression, 2021.

[4] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness, 2020.