A differentiable Mel spectrogram layer for neural networks

John Martinsson[1] and Maria Sandsten[2]

***Abstract*— In this paper an empirical analysis is performed of the differentiable spectrogram as a trainable layer in linear and convolutional neural networks. A Gaussian window, with a window scaling parameter that can be jointly optimized with the neural network through gradient backpropagation, is used in the short-time Fourier transform. The analysis is performed on a theoretically well motivated synthetic classification task suitable for the study of trainable time-frequency transforms. We also derive an expression for the change in concentration of a Gaussian component in a spectrogram for different choices of the window scaling parameter and evaluate the convergence rate to the optimal spectrogram. Finally, a differentiable Mel spectrogram for audio classification is introduced, which is evaluated on the Free Spoken Digits dataset for a linear and a convolutional neural network. The differentiable Mel spectrogram with a trainable parameter achieves a higher test accuracy on average than the standard Mel spectrogram with a fixed parameter for nearly all presented initial values of the parameter on this dataset.**

I. Introduction

The common practice in audio classification using machine learning is to first derive a time-frequency image computed through the short-time Fourier transform (STFT). The squared magnitude of the STFT, the spectrogram, is typically mapped onto the Mel scale using a set of Mel filter banks and then used as input to the neural network model.

An interest for using various forms of trainable time-frequency transforms is seen in classification of audio data, typically bioacoustics [1], human speech, and music and recordings.

Fundamental to these trainable time-frequency transforms is the STFT, and a key hyper parameter for the STFT is the window scaling parameter λ which controls the time-frequency resolution, limited by the uncertainty principle, in the image. Different trade-offs between time and frequency resolution may be optimal for different tasks, and careful consideration of λ has to be done for each task.

Recent work has proposed the differentiable STFT [2], [3] where the parameter λ can be jointly optimized with the neural network. Also, a more general differentiable time-frequency transform called the K-transform[4]. While highly expressive the K-transform is computationally demanding due to the dependence on the Wigner-Ville transform.

In this paper we propose a theoretically well motivated Gaussian-pulse dataset for evaluation of trainable time-frequency transforms, and study the performance of the differentiable spectrogram on this task. In addition, we propose the differentiable Mel spectrogram for audio classification and evaluate it on the Free Spoken Digits (FSD) dataset. The ability to optimize the time-frequency resolution and the machine learning model jointly can improve performance and reduce the needed training computation.

II. Differentiable fixed-size (Mel) spectrogram

The model layer studied in this paper is the differentiable fixed-size spectrogram defined as

$$S_x(t,f) = |F(t,f)|^2 = \\ = |\int_{-\infty}^{\infty} x(s-t)h(s)\exp(-i2\pi fs)ds|^2 \qquad (1)$$

where $F(t,f)$ is the short-time Fourier transform (STFT), $x(t)$ is the signal and

$$h(t) = \exp(-\frac{t^2}{2\lambda^2}) \qquad (2)$$

is a Gaussian window with scaling parameter λ.

The STFT is differentiable with respect to the window parameter λ according to

$$\frac{dF(t,f)}{d\lambda} = \int_{-\infty}^{\infty} x(s-t)\frac{dh(s)}{d\lambda}\exp(-i2\pi fs)ds, \qquad (3)$$

and a differentiable loss function $\mathcal{L}$ is differentiable w.r.t λ through gradient backpropagation using

$$\frac{d\mathcal{L}}{d\lambda} = \sum_{n=1}^{N}\sum_{k=0}^{K-1}\frac{d\mathcal{L}}{dF(n,k)}\frac{dF(n,k)}{d\lambda}, \qquad (4)$$

where $F(t,f)$ is discretized to $F(n,k)$ with a fixed number of N bins in time and K bins in frequency [3].

The differentiable fixed-size Mel spectrogram is an extension proposed in this paper where a set of triangular filter banks are applied to the spectrogram to map it to the Mel scale. This is just a multiplication with a fixed matrix which preserves the gradients during backpropagation.

III. Gaussian-pulse dataset

The Gaussian-pulse dataset defines a classification task consisting of three different classes. Let

$$g(t_0,f_0,\sigma) = A_r\exp(-\frac{(t-t_0)^2}{2\sigma^2})\sin(2\pi f_0 t + \phi_r), \qquad (5)$$

define a Gaussian-pulse with random amplitude $A_r \sim U\{0.5, 1.0\}$ and random phase $\phi_r \sim U\{0, 2\pi\}$, where $U\{a,b\}$ denotes the uniform distribution between a and b. Three classes are simulated. Class 0 consists of a single Gaussian-pulse $x_0(t) = g(t_c, f_c, s_r\sigma) + n(t)$, with time-frequency center (t_c, f_c) and $s_r \sim U\{s_{min}, s_{max}\}$ as a

Thanks to SSF for funding.

[1] J. Martinsson is with RISE Research Institutes of Sweden, Computer Science `john.martinsson@ri.se`

[2] M. Sandsten is with the Centre for Mathematical Sciences, Lund University, Lund, Sweden `maria.standsten@matstat.lu.se`

random scaling factor for the pulse length. The disturbance $n(t)$ is white Gaussian noise with standard deviation $\sigma_n = 0.5$. Class 1 consists of two Gaussian-pulses $x_1(t) = g(t_c - t_r, f_c, \sigma) + g(t_c + t_r, f_c, \sigma) + n(t)$, with the same frequency center f_c and pulse length σ, but with a random displacement $t_r \sim U\{t_{min}, t_{max}\}$ around the time center. Similarly, class 2 consists of two Gaussian-pulses, $x_2(t) = g(t_c, f_c - f_r, \sigma) + g(t_c, f_c + f_r, \sigma) + n(t)$, with the same time center t_c and pulse length σ, but with a random displacement $f_r \sim U\{f_{min}, f_{max}\}$ around the frequency center. The time-frequency center is $(t_c, f_c) = (N/2, K/4)$ and $\sigma = 6.38$ which results in a Gaussian component which is spread across equally many time and frequency bins in the image.

IV. Models

For the Gaussian-pulse dataset we study two models: “LinearNet” and “ConvNet”. The “LinearNet” takes a signal as input (vector), applies the differentiable spectrogram, flattens the derived time-frequency image and applies a linear layer followed by a softmax to the resulting vector. The “ConvNet” the same except for a convolutional layer and a linear layer and ReLU applied after the differentiable spectrogram.

For the FSD dataset we use the same models except for mapping the differentiable spectrogram to the Mel scale (differentiable Mel spectrogram) and log normalization, as is typical for audio classification.

V. Results and discussion

For the “LinearNet” model the non-trainable λ_{init} which on average give the highest test accuracy is 6.38 (the designed optimum for the task), for the “ConvNet” in the range $[5.10, 8.93]$ (see figure 1). We note that both the “LinearNet” and “ConvNet” model with a trainable λ_{init} achieves a higher or similar test accuracy when compared to the non-trainable versions. A model which is robust to the choice of λ_{init} would maximize the area under the curve. Similar results are observed for the “MelLinearNet” and “MelConvNet” models on the FSD dataset when comparing a trainable and non-trainable λ_{init} (see figure 2).

Looking at the converged values of λ_{est} (see figure 1 and 2), we note that for large λ_{init}, with the exception of the “LinearNet” model, the converged λ_{est} is close to the initial value, indicating a weaker learning signal. We perform a theoretical analysis of this phenomenon for Gaussian pulses by deriving the gradient of the area of the pulse (a measure of concentration) in the time-frequency with respect to λ

$$\frac{dA}{d\lambda} = \frac{K}{4\sigma} - \frac{K\sigma}{4\lambda^2}. \tag{6}$$

where K is the FFT-length, and $A = \pi\sigma_t\sigma_f$. We note that for $\lambda \ll \sigma$ a small deviation in λ corresponds to a relatively large change in A compared to if $\lambda \gg \sigma$, where a deviation in λ will cause a smaller change in A. This indicates that for a deviation in λ (coupled to learning rate), the change in concentration and the corresponding resolution of the resulting spectrogram is different for a large initial λ compared to small initial λ.

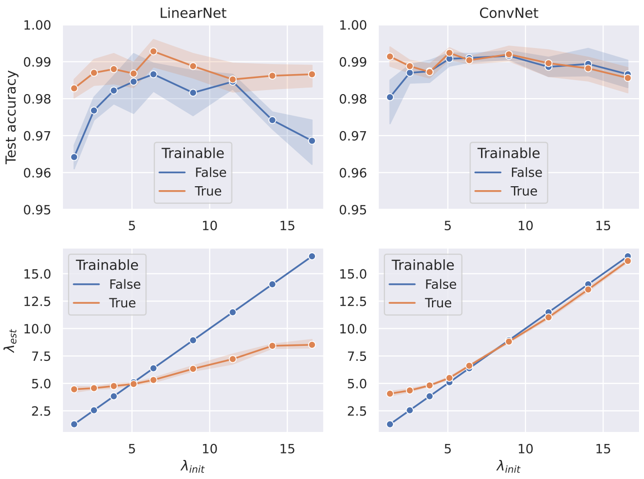

Fig. 1. The average test accuracy and standard deviation of the “LinearNet” model and the “ConvNet” model on the Gaussian-pulse dataset for different initial values of λ (top), and the average estimated window size parameter values λ_{est} for the different initial values (bottom).

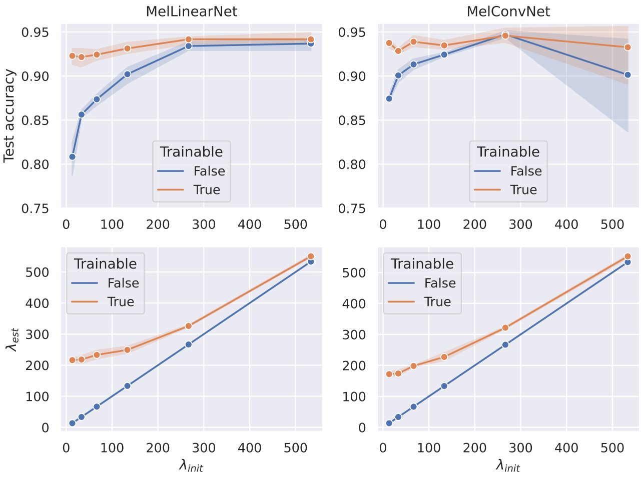

Fig. 2. The average test accuracy and standard deviation of the “MelLinearNet” model and the “MelConvNet” model on the FSD dataset for different initial values of λ (top), and the average estimated window size parameter values λ_{est} for the different initial values (bottom).

References

[1] Mark Anderson and Naomi Harte. Learnable Acoustic Frontends in Bird Activity Detection. In *International Workshop on Acoustic Signal Enhancement, IWAENC 2022 - Proceedings*, 2022.

[2] An Zhao, Krishna Subramani, and Paris Smaragdis. Optimizing short-time Fourier transform parameters via gradient descent. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2021-June(2):736–740, 2021.

[3] Maxime Leiber, Axel Barrau, Yosra Marnissi, and Dany Abboud. A differentiable short-time Fourier transform with respect to the window length. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1392–1396, 2022.

[4] Randall Balestriero, Hervé Glotin, and Richard G Baraniuk. Interpretable and Learnable Super-Resolution Time-Frequency Representation. *Proceedings of Machine Learning Research vol*, 145:1–35, 2021.