Automated Music Transcription

using Convolutional Neural Networks

Cota Ionas-Calin

A thesis presented for the degree of
Bachelor of Computer Science

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ
ΠΣ
ROMÂNIA
BABEȘ-BOLYAI

Computer Science
Babes Bolyai University
Romania
4/20/2019

Contents

List of Figures

List of Tables

1 Theoretical Background

1.1 Introduction to Deep Learning

This section briefly introduces the concept of deep learning.

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine learning algorithms build a mathematical model of sample data, known as ”training data”, in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. The represantation of data fed into a ML algorithm plays a major role, as it affects the algorithm's ability to efficiently extract signals and make decisions. Thus, it is important to carefully select the information included in such a representation. Formally, the representation is com- posed of multiple features extracted from raw data. The process of creating new features requires good and intuitive understanding of the data at hand, becoming incrementally time-consuming with the sophistication of the new features. Thus, the biggest challenge of handcrafted features is deciding which features are important and relevant to the problem [1]

1.2 Neural Networks

This section introduces the main concepts related to neural networks. Neural networks have been around since the 1940s and could initially handle only one hidden layer. But with the development of technologies and hardware it became possible to build deeper, more effective architectures, which leads to deep learning as we know it today.

1.2.1 Brief History

At first, neural networks were inspired by how the biological brain works, which is why deep learning was also called artificial neural networks (ANNs)[1]. In biology, a neuron is the cell that receives, processes and transmits information to other neurons through connections called synapses [2]. On the other hand, artificial neurons are defined as computational units (usually mathematical functions) that take one or more inputs and generate an output.

McCulloch and Pits designed an initial version of the neuron as a linear model in 1943, aiming to replicate brain function [3]:

$$f(x, w) = x1 * w1 + x2 * w2 + ... + xn * wn \tag{1}$$

where $x_1, ..., x_n$ are the input values and $w_1, ..., w_2$ is a set of hand-chosen weights.

1.2.2 Components of an artificial neural network

A simple artificial neural network (ANN) consists of input layer, hidden layer and output layer, where the values of the hidden layer are used as inputs for the output layer. A network with several layers is known as a deep neural network. Data flows through the neurons of the layer. Each neuron transforms the input it receives and sends it to the next layer. The neurons share the same characteristics regardless of the layer they are part of.

The Neuron, also called node, is the basic unit of a neural network. It's main components include inputs, weights, activation function and output(s). From a high level point of view the inputs are multiplied by weights, then an activation function is applied to the result and finally, another function computes the output[4][5].

- Weights are defined as adaptive coefficients, whose values are changed during the learning process. They represent the strength of the connection between units. A weight decides how much impact the input will have on the output.

- The summation function helps combine the input and weights, before passing the result to the activation function. Denote the input as $X = [x_1, x_2, ...x_n]$ and the weight vector as $W = [w_1, w_2, ...w_n]$.
  The summation function could be defined as the dot product between these two vectors:

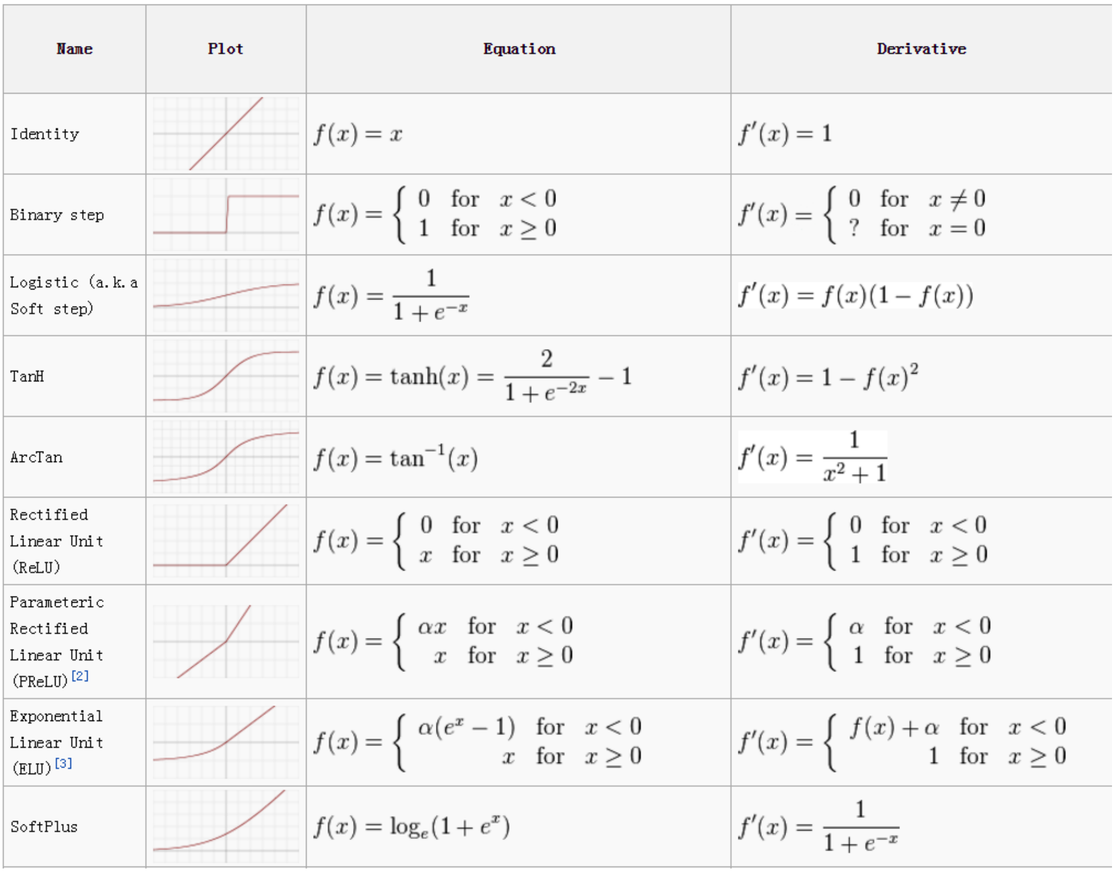
  $$X \cdot W = x_1 \cdot w_1 + x_2 \cdot w_2 + ... + x_n \cdot w_n \tag{2}$$

  The summation function could instead compute the minimum, maximum etc. depending on the designated network architecture.The simplest form of an artificial neuron is a linear function which computes the weighted sum of inputs, to which, optionally, bias can be added:

  $$y = \sum_{i=1}^{i=n} (x_i \cdot w_i) + b, \text{ where b is the bias, } x_i \in X, w_i \in W \tag{3}$$

- The activation function transforms the result of the summation function (usually) in a non-linear way. Typically, it has a squashing effect. It serves as a threshold. It divides the original space into two partitions. It's main purpose is to make the neural network non-linear. We denote the activation function as g.

$$y = g(\sum_{i=1}^{i=n}(x_i \cdot w_i) + b), \text{ where b is the bias, } x_i \in X, w_i \in W \qquad (4)$$



Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for} \quad x \neq 0 \\ ? & \text{for} \quad x = 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1 - f(x))$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU)[2]		$f(x) = \begin{cases} \alpha x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$
Exponential Linear Unit (ELU)[3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$

Figure 1: Common activation functions.
Source: http://prog3.com/sbdm/blog/cyh24/article/details/50593400

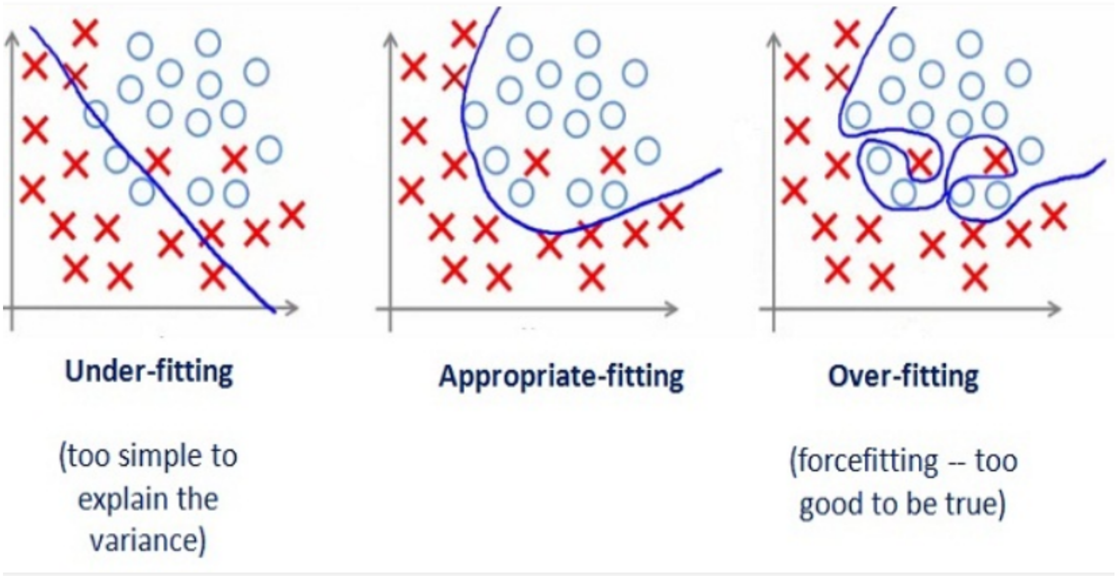
- The output is usually the result of an activation function.

1.2.3 Underfitting and Overfitting

Neural networks are able to learn complicated non-linear functions to fit any training set. On the downside, this may lead to overfitting where the neural network learns the training data so well that it is unable to generalize on new, unseen data. This problem can especially occur on datasets with a small amount of data to learn from.

Underfitting, the counterpart of overfitting, happens when a machine learning model isn't complex enough to accurately capture relationships between a dataset's features and a target variable. An underfitted model results in problematic outcomes on new data, or data that it wasn't trained on, and many times performs poorly even on training data.

Figure 2: Example of overfitting and underfitting
image source vitalflux.com


Under-fitting
(too simple to explain the variance)
Appropriate-fitting
Over-fitting
(forcefitting -- too good to be true)

1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of Deep Neural Networks, specialized in analyzing images. They were inspired by biological processes. The connectivity pattern between neurons resembles the animal visual cortex. [6]

1.3.1 Find subsections

1.4 Sound

Sound is produced when something vibrates. The vibration causes the medium around it to vibrate as well. Vibrations in the air are called traveling longitudinal waves [7], which we can hear. A sound wave is made out of two areas of high and low pressure called compressions and rarefactions (figure 3).

The pattern of the wave repeats after one wavelength. The height of the wave is called Amplitude. It is what determines how load the sound will be, the greater the louder.

The wavelength and the speed of the wave determine the pitch (frequency of the sound).

$$Speed = Frequency \cdot Wavelength \tag{5}$$

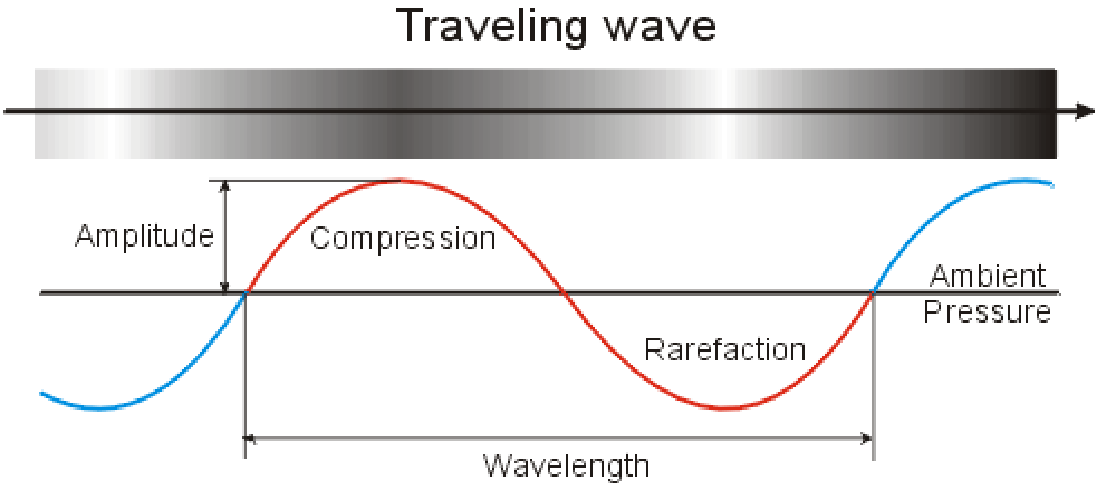


Figure 3: Source: https://method-behind-the-music.com/mechanics/physics/

1.4.1 Pitch

In music, the pitch tells how low or high a note is. In physics, it's measured in a unit called Hertz (Hz) and it's known as frequency. A note that vibrates at 256Hz will be caused by a sound wave vibrating at 256 times/second.

The speed is influenced by the medium in which it travels. Under standard temperature and pressure sound's speed is 343 meters per second. [8]

The equation (5) can be rewritten as:

$$Frequency = \frac{Speed}{Wavelength} \tag{6}$$

1.5 Digital Signal Processing

Digital Signal processing (DSP) is an engineering field focused on analyzing and altering digital signals. It takes real-world signals like voice, audio, video and then mathematically manipulates them. [9]

Signals need to be processed so that the information they contain can be displayed, analyzed or converted to another type of signal. Analog-to-Digital converters take signals from the real-world and turns them into binary digital format. At this point the DSP takes over by capturing the digitized information and processes it, later to be fed back for use in the real-world.

1.5.1 Discrete Fourier Transformation

The Discrete Fourier Transformation (DFT) is one of the most important operation of DSP. It is any quantity or signal that varies over time, such as the pressure of a sound wave, sampled over a finite time interval (often defined by a window function). [10]

$$X[k] = \frac{1}{N}\sum_{j=0}^{N-1}(x[j] \cdot e^{-j \cdot (\frac{2\pi}{N})) \cdot n \cdot k} \quad \text{for k = 0...N-1} \tag{7}$$

The DFT tells you what frequencies are present in your signal and in what proportions.

It has a complexity of $O(n^2)$ so in practice the Fast Fourier Transform (FFT) algorithm is used instead. FFT runs in $O(n \cdot log(n))$

1.6 Music Transcription

1.6.1 Traditional Music Transcription

1.6.2 Automated Music Transcription

2 Related Work

Automatic music transcription (AMT) has been attempted since the 1970s and polyphonic music transcription dates to the 1990s [11]

2.1 State of the art in AMT

A model used in [12] uses 87 Support Vector Machine (SVM) classifiers to perform frame-level classification with the advantage of simplicity, and then a Hidden Markov Model (HMM) post-processing was adopted to smooth the results. On top of it, Deep Belief Network(DBN) was added to learn higher layer representation of features in [13]. Since none of the approaches has reached the same level of accuracy as human experts, most music transcription work is completed by musicians. With the development of deep learning in recent years, many researchers were inspired to apply networks to accomplish AMT. A model based on Convolutional Neural Networks (CNN) was proposed in [14]. More models adopted Reccurent Neural Networks (RNN) or Long Short-Term Memory (LSTM) due to its capability of dealing with sequential data [11] [15] [16]. In [17], 5 models were compared and the ConvNet model was reported as resulting in the best performance.

The first majort AMT work is Smaragdis et al.[18]. This approach uses Non-Negative Matrix Factorization (NMF). This is the main methodology employed in software for automatic transcription, but it has it's limitations. For example, it needs to know how many individual notes are desired for the transcription (information that is not always available).

The next work worth mentioning is Emiya et al.[19], not because of their transcription system (as it was out-performed in the same year), but because of the dataset they created that has become the standard in evaluating any multi-pitch estimation system. They created the MIDI-Aligned Piano Sounds (MAPS) data set composed of around 10,000 piano sounds either recorded by using an upright Disklavier piano or generated by several virtual piano software products based on sampled sounds. The dataset consists of audio and corresponding annotations for isolated sounds, chords, and complete pieces of piano music. For our purpose we use only the isolated sounds (daca nu gasesc ceva mai bun).

Sigtia et al.[20] built the first AMT system using CNN, outperforming the state of the art approaches using NMF. Convolutional Neural Networks are a discriminative approach to AMT, which has been found to be a viable alternative to spectrogram factorization techniques. Discriminative approaches aim to directly classify features extracted from frames of audio to the output pitches. This approach uses complex classifiers that are trained using large amounts of training data to capture the variability in the inputs, instead of constructing an instrument specific model.

2.2 Products

In this section, we present products that use deep learning for AMT.

2.2.1 Melodyne

Melodyne is a popular plugin used for Music Transcription and Pitch Correction. It costs up to $700.

The Melodic and Polyphonic algorithms offer you, in the case of vocals as well as both mono- and polyphonic instruments, full access to the notes of which the sound is composed as well as to their musical parameters.

2.2.2 AnthemScore

AnthemScore is a product used for Music Transcription that uses CNN.

They approach note detection as an image recognition problem by creating spectrograms of the audio. They show how the spectrum or frequency content changes over time. The method used for creating the spectrograms is the constant Q transform instead of the more common Short Time Fourier Transform (STFT) method.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016. http://www.deeplearningbook.org.

[2] “Neuron.” https://en.wikipedia.org/wiki/Neuron.

[3] W. . P. McCulloch, *A logical calculus of the ideas immanent in nervous activity.* W. Bulletin of Mathematical Biophysics (1943) 5: 115, 1943. https://doi.org/10.1007/BF02478259.

[4] B. C. Bangal, *Automatic generation control of interconnected power systems using artificial neural network techniques.* 2009. https://doi.org/10.1007/BF02478259.

[5] *Everything you need to know about Neural Networks* . 2018.

[6] “Convolutional neural networks.” https://en.wikipedia.org/wiki/Convolutional_neural_network.

[7] “Physics of sound.” https://method-behind-the-music.com/mechanics/physics/.

[8] “Speed of sound.” https://en.wikipedia.org/wiki/Speed_of_sound.

[9] “Digital Signal Processing.” https://www.analog.com/en/design-center/landing-pages/001/beginners-guide-to-dsp.html.

[10] G. Sahidullah, Md.; Saha, *A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition.* IEEE Signal Processing Letters. 20 (2): 149–152, 2013.

[11] D. G. Morin, *Deep neural networks for piano music transcription.* 2017.

[12] G. E. Poliner and D. P. Ellis, *A discriminative model for polyphonic piano transcription.* EURASIP Journal on Applied Signal Processing, vol. 2007, no. 1, pp. 154, 2007.

[13] J. N. J. Nam, H. Lee, and M. Slaney, *A classification-based polyphonic piano transcription approach using learned feature representations.* ” Proceedings of the 12th International Society for Music Information Retrieval Conference, pp. 16-180, 2011.

[14] K. Ullrich and E. van der Wel, *Music transcription with convolutional sequence-to-sequence models.* ” International Society for Music Information Retrieval, 2017.

[15] J. F. S. B. L. Sturm, O. Ben-Tal, and I. Korsunova, *Music transcription modelling and composition using deep learning.* arXiv preprint arXiv:1604.08723, 2016.

[16] S. Sigtia, E. Benetos, N. B.-L. T. Weyde, A. S. d'Avila Garcez, and S. Dixon, *A hybrid recurrent neural network for music transcription.* IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2061-2065, 2015.

[17] E. B. S. Sigtia and S. Dixon, *An end-to-end neural network for polyphonic piano music transcription.* IEEE/ACM Trans. Audio Speech Lang. Process., 24, 927–939, 2016.

[18] P. Smaragdis and J. C. Brown., *Non-negative matrix factorization for polyphonic music transcription.* In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on., pages 177–180. IEEE,, 2003.

[19] R. B. V. Emiya and B. David., *Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle.* IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1643–1654, 2010.

[20] J. Sleep, *AUTOMATIC MUSIC TRANSCRIPTION WITH CONVOLUTIONAL NEURAL NETWORKS USING INTUITIVE FILTER SHAPES.* A Thesis presented to the Faculty of California Polytechnic State University, San Luis Obispo, 2017.