1. [Musical Instrument Identification Using Deep Learning Approach](https://www.mdpi.com/1424-8220/22/8/3033)

Published: 15 April 2022

**Data**:

The Slakh dataset was used, which contains 2100 audio tracks with aligned MIDI files, and separate instrument stems along with tagging. From all the available instruments, four were selected for the experiment: bass, drums, guitar, and piano.

**Digital Dat**a:

The examples were then stored using the NumPy format on files that contain mixed signals, instrument references, and vectors of labels to indicate which instruments were used in the mix

**Sampling Size**:

To achieve repeatability of the training results, the whole dataset was a priori divided into three parts, but with the condition that a single audio track cannot be split into each part:

* Training set—116,413 examples;
* Validation set—5970 examples;
* Evaluation set—6983 examples.

**Technique:**

The proposed neural network was implemented using the Keras framework and functional API [45]. The model initially produces MFCC (mel frequency cepstral coefficients) from the raw audio signal using built-in Keras methods [46]. The parameters for those operations are as follows:

* 1024 samples Hamming window length
* 512 samples window step
* 40 MFCC bins.

The training was performed using the Tensorflow framework for the Python language:

* 100 epochs with the mean squared error (MSE) as the loss function.
* a single RTX2070 graphics card with an AMD Ryzen 5 3600 processor and 32 GB of RAM.
* a single epoch is about 8 min using multiprocessing data loading
* batch size of 200.

**Results**:

Of about 200 examples, 61 has bad value of false negative examples. A precision of 93% and an F1 score of 0.93 using a simple convolutional network based on the MFCC. AUC ROC of approximately 0.91 or F1 score of about 0.64.

1. [Automatic Assessment of Tone Quality in Violin Music Performance](https://repositori.upf.edu/bitstream/handle/10230/43229/giraldo_fpsyg_autom.pdf)

**Data**:

Tone qualities for evaluation were chosen using the semantic differential method, in which each tone is measured against bipolar scales with opposing extremes. Two sets of recordings were considered: first, the publicly available data set of recorded sounds from the Good-Sounds project, which included a set of recorded violin sounds with good and bad examples of five of the sound quality dimensions. Second set was obtained recorded examples of the expert-defined tonal semantic dimensions from a professional violinist.

**Digital Data**:

Outliers were eliminated and data filtering was performed using IQR filter with:

* extreme value factor set to 3 times the IQR
* outlier value of 1.5 times the IQR

Feature extraction and selection was used to extract audio descriptors from the recorded audio samples.

**Technique**:

The machine learning algorithms taken into consideration were:

* Linear Regression
* Support Vector Machines (SVM) with radial kernel
  + Sequential Minimal Optimization (SMO) algorithm for training a support vector classifier
* Artificial Neural Networks (ANN) with one hidden layer (half the size of the input nodes).
* Weka machine learning library

The machine learning component of the system was implemented in C++, based on the OpenFrameworks toolkit (for OSC) and the OpenCV library for machine learning and data processing.

1. [Deep Learning-Based Music Quality Analysis Model](https://www.researchgate.net/publication/361280775_Deep_Learning-Based_Music_Quality_Analysis_Model)

**Data**:

The high feature descriptiveness of deep learning requires a huge amount of training data. In the domain of music processing, open source datasets are related to speech, such as music-related open source datasets, e.g., Million song Dataset and MusicNet. There are ambient soundrelated datasets such as the AudioSet.

Data augmentation is used to address the challenge of limited datasets in music quality analysis, the study explores. These include random segmentation and frame skipping to increase data diversity and size, which helps prevent overfitting in deep learning models.

**Technique**:

* MLP: Used for learning global features from one-dimensional vectors like flattened Mel-frequency cepstral coefficients (MFCC).
* CNN: Utilized for extracting local features and learning the relationships between adjacent frames in one-dimensional sound sequences or two-dimensional spectrograms.
* RNN: Employed to model short-term and long-term temporal dependencies in the audio data, although combined with CNNs for better feature extraction and temporal characterization.
* Support Vector Machines (SVM): The research also incorporates SVM classifiers for small-sample music quality evaluation.

1. [Music and Instrument Classification using Deep Learning Technics](https://cs230.stanford.edu/projects_fall_2019/reports/26225883.pdf)

**Data**:

AudioSet by Google provides human labeled data, consisting of a set of 10 second clips from YouTube, labeled with the audio instruments they contain.

2,400 samples for each class; composing a total of 9,600 samples using the same distribution and divided into 8000 samples for training, 800 samples for validation and 800 samples for the evaluation.

**Digital Data**:

Pre-processed the data:

* set the frequency of all audio samples to 16000 Hz
* down mixing them to use 1 channel.
* padded shorter samples to 10 seconds

Leveraged data augmentation techniques, and integrated random white noise to the audio samples of the training set before extracting the mel-spectograms.

**Technique**:

Model used is a model for image classification on the mel-spectrograms. 7-layer CNN, in which each convolution has a kernel size of 2, a stride of 1 and 0 padding.

Trained on 8,000 samples, with batches of size 128, 15 epochs, and a learning rate of 10e-3.

**Results**:

average precision of 70% , an average recall of 65%, and an average F1-score of 64%.

1. [A real-time system for measuring sound goodness in instrumental sounds](https://repositori.upf.edu/bitstream/handle/10230/32131/Romani_aes_rea.pdf?sequence=1&isAllowed=y)

Data: <https://www.upf.edu/web/mtg/good-sounds>

These attributes considered for sound quality are dynamic stability, pitch stability, timbre stability, timbre richness and attack clarity. This work is carried out by first creating a training dataset of single note recordings including six classes of sounds per studied instrument (i.e., clarinet, flute and trumpet)

For the good sound class, musicians were asked to play in a mezzoforte dynamic level, while for the other classes no dynamic restriction was imposed. Five of the classes include examples of note recordings which are intentionally badly played according to the five aforementioned sound attributes. The sixth class includes examples of note recordings which are considered to be well played.

Technique:

Using Essentia, different audio descriptors are selected and used for each one of the sound attributes, e.g. YinFFT algorithm for pitch stability or Tristimulus for timbre stability.

Almost 200 features are used for each sound attribute. Then, one single feature is used to determine the goodness of a single sound attribute. This feature is selected using a machine learning process approach: OneR classifier using the Weka data mining software.

Results:

A table with numbers and text

Description automatically generated with medium confidence

1. [Evaluation of CNN-based Automatic Music Tagging Models](https://arxiv.org/pdf/2006.00751)

The article titled "Evaluation of CNN-based Automatic Music Tagging Models" focuses on assessing various convolutional neural network (CNN) architectures designed for automatic music tagging. This task involves assigning relevant tags (like genres, moods, or instruments) to music tracks based on their audio content.

**Models Evaluated**: The paper evaluates multiple CNN-based models:

* **Fully Convolutional Network (FCN)**: Uses Mel-spectrogram inputs, with max-pooling to increase the receptive field.
* **Musicnn**: Incorporates domain knowledge with vertical and horizontal filters to capture timbral and temporal features.
* **Sample-level CNN**: Processes raw audio waveforms, avoiding assumptions made by Mel-spectrogram-based models.
* **Convolutional Recurrent Neural Network (CRNN)**: Combines CNNs for local feature extraction and RNNs for temporal summarization.
* **Self-attention Model**: Uses the Transformer encoder for temporal summarization, adapted from NLP.
* **Harmonic CNN**: Utilizes harmonically stacked inputs and trainable band-pass filters.
* **Short-chunk CNN**: A simple yet effective CNN model trained on short audio chunks.

**Evaluation Metrics**: The models are evaluated using ROC-AUC and PR-AUC scores, which measure the performance of binary classification tasks, particularly useful for multi-label problems like music tagging.

Conclusion: The Harmonic CNN and Short-chunk CNN models performed particularly well across the board.

1. [Audio Spectrogram Transformer](https://arxiv.org/pdf/2104.01778)

AST is the first model that uses a pure Transformer architecture for audio classification tasks. Unlike traditional models that rely on CNNs to process audio spectrograms, AST leverages the self-attention mechanism to capture both local and global dependencies in the data.

**Architecture:**

* **Patch-based Input**: The audio spectrogram is divided into patches, similar to how images are processed in Vision Transformers (ViT). Each patch is treated as a token, and the Transformer processes these tokens to learn the relationships between different parts of the spectrogram.
* **Self-Attention Mechanism**: AST uses multi-head self-attention to model interactions between all patches, allowing the model to capture long-range dependencies in the audio data.

On AudioSet, AST outperformed previous state-of-the-art models, achieving a significant improvement in classification accuracy. The study highlights that while CNNs are effective in capturing local features in spectrograms, they may struggle with long-range dependencies. AST, with its attention mechanism, overcomes this limitation by considering the entire spectrogram context, leading to improved classification performance.

1. [Attention Is All You Need](https://arxiv.org/pdf/1706.03762)

Introduces the Transformer model, a new architecture for sequence transduction tasks, such as translation, without relying on recurrent or convolutional layers

**Transformer Architecture**:

* The Transformer model uses a self-attention mechanism to draw global dependencies between input and output. It entirely discards the use of recurrence and convolutions.
* The architecture is simpler and more parallelizable compared to RNNs and CNNs, making it faster to train.

**Self-Attention Mechanism**:

* The core component of the Transformer is the self-attention mechanism, which allows the model to focus on different parts of the input sequence when producing each output element.
* Self-attention operates in three steps: it computes three vectors (query, key, and value) for each input token, computes attention scores, and then aggregates the results using these scores.

**Multi-Head Attention**:

* To capture different aspects of the input sequence, the Transformer uses multi-head attention, where multiple attention mechanisms run in parallel, and their outputs are concatenated.
* This allows the model to jointly attend to information from different representation subspaces.

**Performance and Results**:

* The Transformer outperformed state-of-the-art models like RNNs and CNNs in tasks such as translation (e.g., English-to-German and English-to-French), with significant improvements in both speed and accuracy.
* It also scales well, demonstrating the effectiveness of the self-attention mechanism in handling long-range dependencies in sequences.

1. [Evaluation of CNN-based Automatic Music Tagging Models](https://arxiv.org/pdf/2006.00751)

Explores the performance of various Convolutional Neural Network (CNN) models in the context of music tagging—a task where models automatically assign descriptive tags to music tracks based on their audio content.

**Dataset and Evaluation:** The study leverages the MTG-Jamendo dataset, a large-scale music tagging dataset, and evaluates the models based on their ability to correctly predict tags. This dataset includes various music genres and attributes, making it suitable for evaluating model performance across different contexts.

**Performance Metrics:** The authors use several performance metrics, including precision, recall, F1 score, and Area Under the Receiver Operating Characteristic (ROC-AUC) curve, to comprehensively assess the models' effectiveness.

**Key Findings:** The study finds that while deeper CNNs can achieve better performance, the gains may plateau or diminish if the model becomes too complex relative to the data quality.

1. [AST: Audio Spectrogram Transformer](https://arxiv.org/pdf/2104.01778)

It introduces the Audio Spectrogram Transformer (AST), which processes 2D audio spectrograms as visual tokens, leveraging the Vision Transformer (ViT) architecture.

**Transformer-based Model**: The Audio Spectrogram Transformer (AST) adapts the Vision Transformer (ViT) architecture for audio classification tasks.

**Spectrogram Processing**: AST processes 2D audio spectrograms as input, treating them like images.

**Patch Embeddings**: The model uses patch embeddings to capture local and global patterns in audio data.

**State-of-the-art Performance**: AST achieves superior performance on multiple audio classification benchmarks, outperforming traditional convolutional neural networks (CNNs).

**Scalability and Flexibility**: The transformer model's flexibility allows it to be effectively applied to various audio-related tasks.

1. [Learning to Recognize Musical Genre from Audio](https://arxiv.org/pdf/1803.05337)

The paper aims to improve the accuracy of musical genre classification by leveraging deep learning techniques. The focus is on classifying genres directly from raw audio data.

**Dataset:** The authors use the GTZAN dataset, which contains 1,000 audio tracks across 10 genres, to train and evaluate their models.

**Feature Extraction**:

* The study explores various methods for feature extraction, including Mel-frequency cepstral coefficients (MFCCs), which are commonly used in speech and music processing.
* They also experiment with learning features directly from raw audio using convolutional neural networks (CNNs).

**Model Architecture**:

* The authors propose a CNN architecture that takes raw audio waveforms as input, avoiding the need for hand-crafted features.
* The model includes several convolutional and pooling layers, followed by fully connected layers, leading to a softmax output layer for genre classification.

**Performance**:

* The CNN model trained on raw audio waveforms shows competitive performance compared to traditional approaches using hand-crafted features like MFCCs.
* The study highlights that deep learning models can automatically learn effective features from raw audio, reducing the need for manual feature engineering.

1. [END-TO-END LEARNING FOR MUSIC AUDIO TAGGING AT SCALE](https://arxiv.org/pdf/1711.02520)

The paper aims to improve the accuracy of music audio tagging, which involves assigning descriptive tags to music tracks based on their audio content. The tags can describe genres, moods, instruments, etc. The authors propose an end-to-end learning approach, where the model directly learns to map raw audio waveforms to tags without requiring hand-crafted features or intermediate representations.

**Model Architecture**:

* The model uses convolutional neural networks (CNNs) to process the raw audio waveform.
* The architecture consists of multiple convolutional layers followed by global pooling and dense layers, culminating in a sigmoid output layer for multi-label classification.
* The global pooling layer plays a critical role by aggregating features across the entire audio track, allowing the model to learn relevant features at different temporal scales.

**Dataset and Training**:

* The model is trained on a large-scale dataset comprising millions of music tracks, each annotated with multiple tags.
* Data augmentation techniques are employed to improve the model's robustness, such as time-stretching and pitch-shifting.
* The loss function used is the binary cross-entropy, appropriate for multi-label classification tasks.

**Results**:

* The proposed end-to-end model significantly outperforms traditional models that rely on hand-crafted features.
* The paper demonstrates that the model scales effectively to large datasets and can handle the variability and complexity of real-world music audio.

1. [CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR MUSIC CLASSIFICATION](https://arxiv.org/pdf/1609.04243)

Explores the use of Convolutional Neural Networks (CNNs) combined with Recurrent Neural Networks (RNNs) for music classification tasks. The paper investigates how combining CNNs and RNNs can be used effectively for music classification, leveraging CNNs for feature extraction and RNNs for temporal modeling.

**CNNs for Feature Extraction**:

* CNNs are applied to the input music spectrogram to extract spatially local features. These features capture local patterns in time and frequency, which are crucial for identifying musical elements like pitch and timbre.

**RNNs for Temporal Modelling**:

* RNNs, particularly Long Short-Term Memory (LSTM) networks, are used to capture the temporal dependencies in the features extracted by CNNs. This allows the model to account for sequential and temporal relationships in the music data.

**Architecture**:

* The proposed model integrates a CNN as the first layer to extract high-level features from the input music data. These features are then fed into an RNN, which captures the temporal evolution of these features.
* The architecture allows the model to learn both local (via CNN) and global (via RNN) patterns in the music data.

**Results**:

* The model is evaluated on multiple music classification tasks, including genre classification, tagging, and others.
* Experimental results show that the combined CNN-RNN model outperforms traditional models that use only CNNs or RNNs alone. This demonstrates the effectiveness of the hybrid architecture in capturing both spatial and temporal features.

1. [Music Feature Classification Based on Recurrent Neural Networks with Channel Attention Mechanism](https://www.researchgate.net/publication/352348213_Music_Feature_Classification_Based_on_Recurrent_Neural_Networks_with_Channel_Attention_Mechanism)

The study aims to enhance music feature classification by leveraging Recurrent Neural Networks (RNNs) combined with a channel attention mechanism.

**Recurrent Neural Networks (RNNs):** RNNs are utilized for their ability to capture temporal dependencies in music data, which is crucial for understanding musical patterns and features.

**Channel Attention Mechanism:** This mechanism is introduced to improve the model's ability to focus on relevant features by assigning different attention weights to different channels of the input data.

**Architecture**:

* The model integrates RNNs with a channel attention module to effectively extract and classify music features.
* The channel attention module helps in enhancing the discriminative power of the model by emphasizing important features and suppressing less relevant ones.

**Experimental Results**:

* The proposed model is evaluated on standard music datasets.
* The results demonstrate that incorporating channel attention significantly improves classification performance compared to traditional RNN approaches without attention mechanisms.

1. [SSAST: Self-Supervised Audio Spectrogram Transformer](https://arxiv.org/pdf/2110.09784)

The paper introduces SSAST, a self-supervised model for learning audio representations from spectrograms. The aim is to leverage self-supervised learning to improve audio understanding without relying on extensive labeled data.

**Pre-training Task**: The model is trained with a self-supervised pre-training task where the goal is to predict masked parts of the audio spectrogram. This approach helps the model learn useful audio representations by reconstructing missing parts of the input data.

SSAST uses a transformer-based architecture to process audio spectrograms. The model incorporates self-attention mechanisms to capture long-range dependencies in the audio data.

SSAST is shown to be effective for various audio-related tasks, including sound event detection and audio classification, illustrating its versatility and potential in different audio processing scenarios.

1. [Advanced Attention Mechanisms for Long Sequence Transformers](https://medium.com/autonomous-agents/advanced-attention-mechanisms-for-long-sequence-transformers-6c88b2b41514)

The article discusses techniques to improve the efficiency of transformers when handling long sequences.

**Problem:** Standard transformers struggle with long sequences due to quadratic complexity in terms of sequence length, making them inefficient for processing extended inputs.

**Efficient Attention Variants**:

* **Linformer**: Uses low-rank approximations to reduce the complexity of attention operations.
* **Longformer**: Introduces a sliding window approach and global attention to handle longer sequences efficiently.
* **Reformer**: Employs locality-sensitive hashing and reversible layers to reduce computational complexity.

**Transformer Architectures**:

* **Performer**: Utilizes kernel-based approximations to transform attention into linear time complexity.
* **Synthesizer**: Replaces the traditional attention mechanism with parameterized attention matrices, reducing the need for pairwise token interactions.

**Hybrid Models**: Combining various attention mechanisms or integrating them with other techniques to leverage the strengths of each approach while mitigating their weaknesses.

1. [SpecTNT: A TIME-FREQUENCY TRANSFORMER FOR MUSIC AUDIO](https://arxiv.org/pdf/2110.09127)

The paper introduces SpecTNT, a novel transformer architecture designed for analyzing and processing music audio by leveraging time-frequency representations.

SpecTNT operates on time-frequency spectrograms, which capture both temporal and frequency information of audio signals. This approach allows the model to handle the intricate patterns and structures present in music.

The SpecTNT model incorporates several key components:

* Time-Frequency Encoding: Utilizes time-frequency encoding to capture the characteristics of audio signals.
* Self-Attention Mechanism: Employs a self-attention mechanism to capture dependencies within the time-frequency space.
* Positional Encoding: Includes positional encodings to retain the sequential nature of audio data.

1. [Automatic Music Transcription using Convolutional Neural Networks and Constant-Q transform](https://ceur-ws.org/Vol-3486/176.pdf)

**Constant-Q Transform (CQT)**: The CQT is used for time-frequency analysis of audio signals. It is particularly well-suited for music because it provides a logarithmic frequency resolution, which aligns with musical pitch perception.

**Convolutional Neural Networks (CNNs)**: The CNNs are employed to process the time-frequency representations (spectrograms) generated by the CQT. The CNN architecture is designed to learn patterns in these spectrograms that correspond to musical notes.

**Data and Evaluation**: The method was evaluated using a dataset of polyphonic music recordings. The performance was measured in terms of transcription accuracy, including note detection and pitch estimation.

**Results:** The proposed method showed improvements in transcription accuracy compared to traditional methods. It effectively handles polyphonic music and demonstrates the potential of CNNs combined with CQT for automatic music transcription.

1. [Machine Learning Techniques in Automatic Music Transcription: A Systematic Survey](https://arxiv.org/html/2406.15249v1)

The paper introduces AMT as the process of converting audio recordings into symbolic music representations, such as sheet music or MIDI files. It highlights the challenges involved, including pitch detection, rhythm extraction, and handling complex musical elements.

**Machine Learning Techniques:**

* **Supervised Learning:** Techniques that rely on labeled training data, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).
* **Unsupervised Learning:** Methods that do not require labeled data, including clustering and dimensionality reduction techniques.
* **Semi-Supervised and Self-Supervised Learning:** Approaches that leverage a combination of labeled and unlabeled data or use self-generated labels for training.

**Types of Music Transcription:**

* **Frame-level transcription:** also referred to as Multi-Pitch Estimation (MPE), estimates the number and pitch of notes concurrently present in each time frame, typically with a latency of approximately ten milliseconds
* **Note-Level Transcription in Polyphonic Music**
* **Stream-level transcription**: also known as Multipitch streaming (MPS), is a technique that groups estimated pitches or notes into streams, where each stream typically representing an individual instrument or musical voice. This technique is closely related to instrument source separation.
* **Notation-level transcription:** aims to convert audio files into a human-readable musical score, necessitating a comprehensive grasp of musical structures such as harmonic, rhythmic, and stream structures.

**Dataset:**

The MAESTRO (“MIDI and Audio Edited for Synchronous Tracks and Organization”) dataset, contains over a week of paired audio and MIDI recordings from nine years of International Piano-e-Competition events. The dataset includes annotation of isolated notes, duration of notes, chords, and complete piano pieces. The MIDI data includes key strikes, velocities and sustain pedal positions.

1. [Analysis of Piano Performance Characteristics by Deep Learning and Artificial Intelligence and Its Application in Piano Teaching](https://www.researchgate.net/publication/358182204_Analysis_of_Piano_Performance_Characteristics_by_Deep_Learning_and_Artificial_Intelligence_and_Its_Application_in_Piano_Teaching)

The paper discusses the integration of deep learning (DL) and artificial intelligence (AI) in piano teaching. It highlights the use of a convolutional neural network (CNN) for detecting piano note onset, which enhances the learning experience for preschool children aged 4-6.

**Detection Method:** A novel method for detecting piano note onset using convolutional neural networks (CNNs) is presented, showcasing its effectiveness in analyzing musical cues.

**CNN Performance:** The CNN model demonstrated good performance in detecting note onset, with results indicating reduced loss value over training iterations.

The study found that intelligent piano lessons significantly increased children's interest in learning and were well-received by both students and their parents.