Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.678

REVIEW ARTICLE

Cross-domain synergy: Leveraging image processing techniques for enhanced sound classification through spectrogram analysis using CNNs

Valentina Franzoni[1,2]

[1] *Department of Mathematics and Computer Science, University of Perugia, 06123 Perugia, Italy; valentina.franzoni@unipg.it*

[2] *Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*

ABSTRACT

In this paper, the innovative approach to sound classification by exploiting the potential of image processing techniques applied to spectrogram representations of audio signals is reviewed. This study shows the effectiveness of incorporating well-established image processing methodologies, such as filtering, segmentation, and pattern recognition, to enhance the feature extraction and classification performance of audio signals when transformed into spectrograms. An overview is provided of the mathematical methods shared by both image and spectrogram-based audio processing, focusing on the commonalities between the two domains in terms of the underlying principles, techniques, and algorithms. The proposed methodology leverages in particular the power of convolutional neural networks (CNNs) to extract and classify time-frequency features from spectrograms, capitalizing on the advantages of their hierarchical feature learning and robustness to translation and scale variations. Other deep-learning networks and advanced techniques are suggested during the analysis. We discuss the benefits and limitations of transforming audio signals into spectrograms, including human interpretability, compatibility with image processing techniques, and flexibility in time-frequency resolution. By bridging the gap between image processing and audio processing, spectrogram-based audio deep learning gives a deeper perspective on sound classification, offering fundamental insights that serve as a foundation for interdisciplinary research and applications in both domains.

Keywords: audio processing; time-frequency representation; feature extraction; convolutional neural networks; segmentation; pattern recognition; filtering; spectrogram analysis; interdisciplinary research

ARTICLE INFO

Received: 1 June 2023
Accepted: 28 July 2023
Available online: 28 August 2023

1. Introduction

Convolutional neural networks (CNNs)[1–3] often achieve better performance on spectrograms compared to other deep learning methods applied directly to raw audio data. In fact, CNN-based deep learning can benefit from a large corpus of research, to set up and fine-tune the network for the scope of image analysis, which does not happen for every other deep neural network technique. Moreover, we can both exploit traditional machine-learning algorithms on numerical data extracted from spectrograms both CNN-based learning using advanced techniques, e.g., transfer learning. Transfer learning is a method for reusing a pre-trained neural network model for a new task by transferring knowledge from the original task to the new one. This is especially helpful when working with small datasets because it enables the model to draw on knowledge learned from larger datasets.

Transfer learning can be used in the setting of CNN-based deep learning of spectrograms by using a pre-trained CNN model as a

feature extractor[4]. The pre-trained CNN, using the information of a huge number of sample images for training, will have the capability to well recognize basic low-level features of images. This model is used to extract relevant features from the spectrograms, which are then fed into a newly trained classifier to accomplish the desired task[5]. Using this method, we can significantly reduce the amount of training data needed while improving model performance. There are several stages involved in applying transfer learning to CNN-based deep learning of spectrograms. To begin, a pre-trained CNN model is chosen (e.g., ImageNet[6], Inception Resnet[7], GoogleNet[8]), and its convolutional layers are frozen, which means that their weights are not changed during training. The original CNN model's completely connected layers are then removed and replaced with new fully connected layers tailored to the new task. Intermediate layers before the last one can be also added for specific reasons. Backpropagation is then used to train the new completely connected layers on the new task while the weights of the frozen convolutional layers remain constant.

Overall, transfer learning can be a potent tool for improving the performance of CNN-based deep learning models when applied to small datasets (e.g., spectrograms for a particular task or event) which, as images, will share similar low-level features to those used to train the pre-trained model. In this way, image-based deep learning can reach with relative ease good performance. Apart from the benefits provided by pre-trained models, other advantages of transforming audio signals into spectrograms include the interpretability of the data by humans, and the possibility to build hybrid models combining the output from deep neural networks and traditional machine learning applied on numerical data extracted from the spectrogram. In the following subsections, some of these benefits are provided with explanations.

1.1. Localized feature detection

Convolutional neural networks are designed to exploit the local structure and spatial relationships in data, making them well-suited for analyzing spectrograms, which are essentially two-dimensional visual representations of audio signals. In a spectrogram, local patterns correspond to specific audio events or properties, such as harmonic structures, transients, or spectral shapes. CNNs can effectively learn these localized features through convolutional layers, pooling layers, and hierarchical feature extraction, even better if transfer learning is used with models pre-trained on large image datasets.

1.2. Reduced data complexity

Converting raw audio data into spectrograms reduces the complexity of the data by representing it in a more compact and structured format. Spectrograms emphasize the time-frequency characteristics of audio signals, making it easier for CNNs to identify and learn relevant features. This transformation also helps mitigate the challenges posed by the high dimensionality and variability of raw audio data, which can make learning difficult for other deep learning methods.

1.3. Invariance to translation and scale

CNNs are inherently robust to translation and scale variations in the input data due to the shared weights and pooling operations in the convolutional layers[9]. This property is advantageous when working with spectrograms, as it allows the model to recognize patterns and features regardless of their position or scale in the time-frequency representation.

1.4. Transfer learning

As introduced, many pre-trained CNN models are available for image classification tasks, and these models can be fine-tuned for spectrogram-based audio classification tasks by leveraging transfer learning. This approach can lead to significant improvements in performance and reduce training time compared to training a model from scratch.

1.5. Compatibility with image processing techniques

By representing audio signals as spectrograms, we can leverage a wide range of image processing techniques for tasks such as noise reduction, filtering, segmentation, and feature extraction[4]. These techniques can help enhance the quality of the spectrograms and improve the performance of the subsequent pattern recognition or classification tasks.

1.6. Flexibility in time-frequency resolution

Spectrograms offer flexibility in adjusting the time-frequency resolution, which can be tailored to specific applications or requirements. For example, short-time Fourier transform (STFT) spectrograms[10] provide a balance between time and frequency resolution, while Constant-Q transform (CQT) spectrograms[11] offer better frequency resolution at lower frequencies and better time resolution at higher frequencies. This flexibility allows for better analysis and representation of the audio signals depending on the task at hand, choosing each time the more suitable spectrogram representation.

1.7. Human interpretability

Spectrograms provide a visually intuitive representation of audio signals, making it easier to identify and understand patterns, structures, and features in the data[12]. This interpretability can be useful for debugging, feature engineering, and model evaluation. For instance, a practical method to use audio for classification is to extract it from videos which meaning in terms of class labeling is clearly given by the video images. Using long sequences of video and audio frames could lead to including in our data some noise or outlier audio overlapping from other sources, as well as parts where audio is absent: identifying such frames would need a significant human effort, listening every single audio section and editing the audio files. Transforming the audio sections in images with a proper spectrogram model allows to quickly identify by sight silence, white noise, and other outliers that should be discarded from the data set. In some cases, also the different classes can be identified by sight, and a human analysis of eventual classification errors can be exploited for a deeper comprehension of the mistakes induced by each step of the implementation of a classification algorithm.

2. Processing techniques shared between image and audio data analysis

Besides the use of spectrograms, image and audio processing share some processing techniques.

In this section, a comprehensive description is given of the main parallels between image and audio processing techniques, with a focus on spectrograms. In the following section 3, each of these techniques will be seen in more depth with a parallel between the analysis of images and audio data as spectrograms.

2.1. Feature extraction

In both image and audio processing, extracting meaningful features is crucial. For images, this might include edge detection, texture analysis, or shape descriptors[4]. For spectrograms, features such as spectral peaks, energy distribution, and local patterns can be extracted. In both cases, mathematical techniques like Fourier analysis[10], wavelets[13], and Gabor filters[14] can be employed. In fact, both image processing and audio processing involve extracting meaningful features from data, and some mathematical techniques are applicable to both domains. For instance, Fourier analysis can be used to analyze the frequency content of images or spectrograms, while wavelets and Gabor filters can be employed for texture analysis in images or to analyze the time-frequency content of spectrograms.

2.2. Dimensionality reduction

Both image and audio processing often involve reducing the dimensionality of the data for more efficient processing and analysis. Techniques such as principal component analysis (PCA)[15], independent component

analysis (ICA)[16], and non-negative matrix factorization (NMF)[17] can be applied to both image data and spectrogram data.

2.3. Noise reduction and filtering

Images and spectrograms both suffer from the presence of noise, which can interfere with the analysis and understanding of the data. Common mathematical techniques like gaussian filters[18], median filters[19], and adaptive filters[20] can be applied to both types of data to reduce noise and enhance signal quality.

2.4. Segmentation

In image processing, segmentation refers to the partitioning of an image into meaningful regions or objects. Similarly, in audio processing, segmentation can involve dividing a spectrogram into time-frequency regions representing distinct audio events or sources. Techniques like thresholding[21], region growing[22], and graph-based[23] methods can be applied to both image and spectrogram segmentation tasks.

2.5. Pattern recognition and machine learning

Both image and audio processing often involve the use of machine learning algorithms besides CNNs, e.g., to learn and recognize patterns in the data. As introduced, CNNs have been proven effective in tasks like image classification, object detection, and semantic segmentation, as well as in spectrogram-based audio classification and source separation.

3. Application of processing techniques to image processing and audio processing through spectrograms

In this section, we are going to examine each processing technique presented in the previous section 2 in greater detail, with a comparison to the analysis of images and audio data as spectrograms. The particularities of general image analysis and spectrogram analysis are described, providing some example techniques to apply to each different case, and highlighting commonalities or differences in each processing technique.

3.1. Feature extraction

Feature extraction in image processing: Feature extraction involves identifying and describing significant attributes in an image that can be used for further processing, such as classification, segmentation, or object recognition. Some common feature extraction methods used in image processing include:

a. Edge detection: Identifying sharp changes in intensity or color that correspond to object boundaries. Techniques include Sobel, Canny, Prewitt, and Laplacian of Gaussian (LoG) filters.

b. Texture analysis: Describing the spatial distribution of intensity or color values, which can help differentiate between different materials or surface properties. Methods include Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Gabor filters.

c. Shape descriptors: Quantifying the geometric properties of objects or regions in an image, such as the size, perimeter, or compactness. Examples include Hu moments, Zernike moments, and Fourier shape descriptors.

Feature extraction in audio processing with spectrograms: In the context of audio processing, feature extraction involves extracting meaningful attributes from the time-frequency representation of audio signals (i.e., spectrograms). Some common feature extraction methods used with spectrograms include:

a. Spectral features: Quantifying the distribution of energy across different frequency bands. Examples include spectral centroid, spectral bandwidth, spectral rolloff, and spectral contrast.

b. Temporal features: Capturing the variation of audio properties overtime, such as energy, zero-crossing rate, or spectral flux.

c. Harmonic/percussive features: Separating harmonic (tonal) and percussive (rhythmic) components in the spectrogram. Techniques include median filtering and non-negative matrix factorization (NMF).

d. Mel-frequency cepstral coefficients (MFCCs): Describing the shape of the power spectrum of a sound signal using a compact set of coefficients, often used in speech and music processing.

e. Chroma features: Representing the distribution of energy across different musical pitches or chroma values, useful for tasks such as chord recognition or music genre classification.

3.2. Dimensionality reduction

Dimensionality reduction applies a process of reducing the number of variables or features while retaining most of the original data's structure and information. This is important for reducing computational complexity, removing noise, and improving the performance of machine learning algorithms.

Dimensionality reduction techniques can be used to project the data into a lower-dimensional space that captures the most significant patterns or structures. In both image processing and audio processing, dimensionality reduction can be used for visualization, feature extraction, or pattern recognition tasks by revealing relationships and structures in the data.

Some common dimensionality reduction techniques applicable to both image data and spectrogram data include the following.

a. Principal component analysis (PCA): PCA is a linear dimensionality reduction technique that transforms the original data into a new coordinate system such that the greatest variance lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

In image processing, PCA can be used for tasks such as face recognition, where it's known as Eigenfaces.

In audio processing, PCA can be applied to spectrograms to capture the most significant time-frequency patterns and reduce data dimensionality for efficient processing.

b. Independent component analysis (ICA): ICA is another linear dimensionality reduction method that aims to find statistically independent components in the data. Unlike PCA, which focuses on the variance, ICA considers higher-order statistics to find independent sources in the data.

In image processing, ICA can be used for tasks like blind source separation or feature extraction.

In audio processing, ICA can be applied to spectrograms for tasks such as audio source separation or denoising, by finding independent audio sources within the spectrogram data.

c. Non-negative matrix factorization (NMF): NMF is a dimensionality reduction technique that decomposes a non-negative data matrix into the product of two lower-dimensional non-negative matrices. This results in a parts-based, sparse representation of the original data.

In image processing, NMF can be used for tasks like object recognition or image segmentation, where it can help identify meaningful patterns or structures in the data.

In audio processing, NMF can be applied to spectrograms for tasks such as source separation, transcription, or feature extraction, by finding meaningful time-frequency components that represent the underlying audio sources.

d. t-distributed stochastic neighbor embedding (t-SNE): t-SNE is a non-linear dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data in two or three dimensions. It works by minimizing the divergence between two probability distributions: one in the high-dimensional space and one in the low-dimensional space.

In image processing, t-SNE can be used for visualizing clusters or structures in image data, such as in image retrieval or classification tasks.

In audio processing, t-SNE can be applied to spectrogram data for visualizing relationships between different audio samples, aiding in tasks like clustering, classification, or similarity analysis.

3.3. Noise-reduction and filtering techniques

Noise-reduction and filtering techniques are essential for enhancing the quality of images and spectrograms, as noise can interfere with the analysis, understanding, and interpretation of the data. In both image processing and audio processing, the primary goal of noise reduction and filtering techniques is to improve the quality of the data by removing noise and unwanted artifacts while preserving important features and structures. This enhances the performance of subsequent processing steps, such as segmentation, feature extraction, and pattern recognition.

Many noise-reduction and filtering techniques used in both image processing and audio processing share similar mathematical foundations. For example, linear filters like gaussian filters and adaptive filters such as wiener filters rely on convolution operations and statistical properties of the data. Wavelet-based denoising techniques involve the use of wavelet transforms and thresholding operations in both domains.

Both image processing and audio processing employ linear and non-linear filters, as well as adaptive filters that adjust their behavior based on the local characteristics of the data. This highlights the versatility and adaptability of filtering techniques to handle various types of noise and artifacts in different situations.

The application of noise reduction and filtering techniques to both images and spectrograms also highlights the similarities in how the signal is represented and processed in both domains. Images and spectrograms can both be thought of as two-dimensional grids of pixel values or time-frequency bins, and filtering techniques can be applied to these grids to enhance the quality of the data.

Finally, both image processing and audio processing face similar challenges when it comes to noise reduction and filtering, such as preserving important features and structures while removing noise, adapting to varying noise properties across the data, and balancing the trade-off between noise removal and distortion of the original signal.

Some common noise reduction and filtering techniques that can be applied to both image data and spectrogram data include the following.

a. Gaussian filters: Gaussian filters are linear filters that smooth an image or spectrogram by convolving it with a gaussian function. The gaussian function is characterized by its standard deviation (σ), which determines the amount of smoothing.

In image processing, gaussian filters can be used to reduce noise or blur images.

In audio processing, gaussian filters can be applied to spectrograms to smooth time-frequency patterns, which can help reduce noise and enhance signal quality.

b. Median filters: Median filters are non-linear filters that replace each pixel in an image or each time-frequency bin in a spectrogram with the median value of its neighboring pixels or bins. Median filters are particularly effective at removing salt-and-pepper noise or impulsive noise while preserving edges and sharp features.

In image processing, median filters can be used to remove noise while retaining important details in the image.

In audio processing, median filters can be applied to spectrograms to remove isolated noisy time-frequency bins, improving the overall quality of the spectrogram.

c. Adaptive filters: Adaptive filters are filters that change their behavior based on the local characteristics of the data, making them better suited for situations where the noise properties vary across the image or spectrogram. Examples of adaptive filters include the wiener filter, which adjusts its smoothing based on the local signal-to-noise ratio, and the bilateral filter, which combines domain and range filtering to preserve edges while smoothing homogeneous regions.

In image processing, adaptive filters can be used to remove noise while preserving important details and structures in the image.

In audio processing, adaptive filters can be applied to spectrograms to remove noise and enhance signal quality in a context-dependent manner.

d. Wavelet-based denoising: Wavelet-based denoising involves decomposing an image or spectrogram into a set of wavelet coefficients, which represent the data at different scales and resolutions. Noise can be reduced by thresholding or shrinking the wavelet coefficients, followed by reconstructing the denoised image or spectrogram using the inverse wavelet transform. Wavelet-based denoising can be applied to both image data and spectrogram data for noise reduction while preserving important features and structures in the data.

3.4. Segmentation

Segmentation refers to the process of dividing data into meaningful regions, objects, or components. In both image processing and audio processing, segmentation techniques can help identify and separate important structures or elements for further analysis.

Many shared mathematical methods and underlying principles between image processing and spectrogram-based audio processing segmentation techniques can be detected.

Segmentation techniques, such as region growing and clustering, rely on similarity measures to compare pixels or time-frequency bins based on their properties (e.g., intensity, color, spectral content). These measures can include Euclidean distance, cosine similarity, or correlation coefficients. Both image and audio processing use similarity measures to group data points and identify meaningful segments.

Graph-based segmentation methods represent the data as a graph, where nodes correspond to pixels in an image or time-frequency bins in a spectrogram, and edges represent relationships or similarities between them. The use of graph theory in both domains allows for the application of similar algorithms, such as normalized cuts or minimum spanning trees, to identify segments or connected components based on connectivity or similarity.

Segmentation techniques often involve the use of statistical methods to analyze the data and determine optimal thresholds or cluster centroids. For example, thresholding techniques like Otsu's method rely on maximizing the between-class variance, while clustering algorithms like k-means involve minimizing the within-cluster sum of squared distances. Both image and audio processing use these statistical methods to segment data into meaningful regions.

Morphological operations, such as erosion, dilation, opening, and closing, can be used to refine segmentation results in both image and audio processing. These operations involve the use of structuring elements to modify the shape and connectivity of segmented regions. In both domains, morphological operations can help remove noise, fill gaps, and smooth the boundaries of segmented regions.

In this subsection, some shared segmentation techniques and their applications in both image processing and audio processing are analyzed.

a. Thresholding: Thresholding is a simple and widely-used segmentation technique that involves setting a threshold value to separate the data into two or more classes.

In image processing, thresholding can be used to separate objects from the background or to identify regions of interest in an image.

In audio processing, thresholding can be applied to spectrograms to separate time-frequency regions corresponding to different audio events or sources.

b. Region growing: Region growing is a segmentation technique that starts with a seed point (or multiple seed points) and iteratively expands the region(s) by adding neighboring pixels or time-frequency bins that meet a certain similarity criterion.

In image processing, region growing can be used to segment connected objects or homogeneous regions.

In audio processing, region growing can be applied to spectrograms to identify contiguous time-frequency regions representing distinct audio events or sources.

c. Clustering: Clustering is a segmentation technique that groups data points based on their similarity or proximity in feature space. Common clustering algorithms include k-means, hierarchical clustering, and DBSCAN.

In image processing, clustering can be used to segment images into regions with similar color, texture, or other features.

In audio processing, clustering can be applied to spectrograms to group time-frequency regions or components based on their spectral or temporal properties, which can help separate different audio events or sources.

d. Graph-based methods: Graph-based segmentation techniques represent the data as a graph, where nodes correspond to pixels or time-frequency bins, and edges represent the relationships or similarities between them. Examples of graph-based segmentation methods include normalized cuts, minimum spanning tree-based algorithms, and Markov random fields.

In image processing, graph-based methods can be used to segment images into regions or objects based on their connectivity or similarity in feature space.

In audio processing, graph-based methods can be applied to spectrograms to identify connected components or clusters that represent different audio events or sources.

e. Deep learning-based methods: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been widely used for segmentation tasks in both image and audio processing. These algorithms learn hierarchical feature representations from the data and can be trained to segment and classify data points based on their features. In both domains, deep learning methods have demonstrated significant success in segmentation tasks, such as semantic segmentation in image processing and source separation or onset detection in audio processing.

In image processing, deep learning-based methods like U-Net or Mask RCNN can be used for semantic segmentation or instance segmentation tasks.

In audio processing, deep learning-based methods can be applied to spectrograms for tasks like source separation, onset detection, or instrument recognition by learning to segment and classify time-frequency regions or components.

3.5. Pattern recognition and its applications

Pattern recognition involves the identification and classification of patterns, structures, or features in data. In both image processing and audio processing, pattern recognition techniques can help extract meaningful information from the data for various applications. Common pattern recognition techniques and their applications in both image processing and audio processing include the following.

a. Template matching: Template matching is a pattern recognition technique that involves comparing a template or pattern with regions of the data to find matches.

In image processing, template matching can be used for object detection or feature extraction, such as detecting specific shapes or patterns within an image.

In audio processing, template matching can be applied to spectrograms for tasks like onset detection, pitch estimation, or chord recognition by comparing time-frequency patterns with predefined templates.

b. Feature extraction and classification: Feature extraction involves transforming the data into a lower-dimensional feature space, which can be used to classify or recognize patterns.

In image processing, features such as edges, corners, or texture can be extracted and used to classify objects or scenes.

In audio processing, features, e.g., as spectral flux, chroma, or mel-frequency cepstral coefficients (MFCCs) can be extracted from spectrograms and used for tasks like instrument recognition, genre classification, or speaker identification.

c. Machine learning algorithms: Machine learning algorithms, such as support vector machines (SVMs), decision trees, or k-nearest neighbors (k-NN), can be used for pattern recognition tasks in both image and audio processing. These algorithms can be trained on features extracted from the data to learn relationships and classify patterns.

In image processing, machine learning algorithms can be used for tasks like object recognition or scene classification.

In audio processing, machine learning algorithms can be applied to features extracted from spectrograms for tasks like audio event detection, emotion recognition, or music classification.

d. Deep learning algorithms: Deep learning algorithms, such as CNNs and RNNs, have been widely used for pattern recognition tasks in both image and audio processing. These algorithms learn hierarchical feature representations from the data and can be trained to recognize complex patterns.

In image processing, deep learning algorithms like CNNs have demonstrated significant success in tasks like object recognition, image classification, and semantic segmentation.

In audio processing, deep learning algorithms can be applied to spectrograms for tasks like audio tagging, source separation, or speech recognition by learning to recognize time-frequency patterns.

4. Limitations of the use of spectrograms and solutions

The use of spectrograms for audio recognition based on deep learning using CNNs has for sure some limitations. In this section we highlight the limits with their drawback and propose possible mitigations, solutions or alternative techniques.

4.1. Simultaneous sounds

In spectrograms, multiple sound events can sum together, leading to overlapping frequency components that are challenging to separate[24]. The magnitude of a particular observed frequency could be produced by any number of accumulated sounds or complex interactions between sound waves, such as phase cancellation.

The trivial drawback is that overlapping sound events eventually represented in the same spectrogram could cause confusion in the classification process, leading to inaccurate recognition results and reduced performance. Advanced source separation techniques, such as Blind Source Separation (BSS) algorithms[25] or deep learning-based methods specific for the task, could be used to isolate individual sound events from the spectrogram or to recognize their patterns in the mixed representation, making the classification task more

manageable and improving the accuracy of CNN-based sound recognition. For instance, if a CNN is trained to recognize two different sound patterns specific for two classes, its capabilities may still be kept even giving as input the two classes together in the same spectrogram, given that one of them is prevalent and can be recognized as the main predicted label (or both can be recognized if the problem is multiclass). However, the success of recognition in such scenarios depends on the nature of the sounds and the degree of overlap between their spectrogram representations.

4.2. Spatial invariance

Spectrograms have two dimensions representing different units: time and frequency. While horizontal shifts in spectrograms correspond to time offsets, vertical shifts change the meaning of the sound by altering its frequency. This challenges the spatial invariance property of 2D CNNs, which are better suited for tasks where shifts in the spatial domain do not change the semantics. Spatial invariance limitations may result in suboptimal performance of CNNs when classifying sound events represented in spectrograms.

The solution could be to implement 1D CNNs or combining 2D and 1D CNN architectures to effectively capture both temporal and frequency characteristics, allowing the model to recognize patterns over time and frequency without being hindered by spatial invariance issues. Alternatively, wavelet transform[26], which provides localized time-frequency representations, can be explored as an alternative to spectrograms. In a single data set, normalizing the sounds can also partially mitigate this issue, when the particular sound event does not greatly vary across the time.

4.3. Non-locally distributed frequencies

Frequencies in spectrograms, especially for periodic sounds with harmonics, are non-locally distributed. Harmonics are spaced apart based on relationships dictated by the sound source, making their identification challenging for standard local feature extraction methods. The drawback is that non-local frequency distributions pose difficulties in accurately identifying individual harmonics, impacting the CNN's ability to discern sound characteristics effectively.

Incorporating attention mechanisms[27] or long short-term memory (LSTM) layers[28] into the CNN architecture can allow the model to capture temporal dependencies and learn to recognize non-locally distributed frequency patterns, improving the performance of sound recognition from spectrograms. Alternatively, wavelet transform's multi-resolution analysis can provide localized frequency information and enhance harmonic analysis.

4.4. Temporal nature of sound and lack of parallel information

Sound is highly serial, experienced moment by moment, and lacks the parallel information found in images. Sound events unfold over time, and understanding their meaning requires considering the temporal relationships between spectral developments. CNNs are designed to process static images with parallel information and may struggle to effectively handle the temporal nature of sound events represented in spectrograms. Incorporating recurrent neural networks (RNNs)[28] or attention mechanisms[27] that capture temporal dependencies and sequential patterns can enhance the CNN's capability to recognize sound events from spectrograms and leverage their temporal characteristics for more accurate classification, even including sound direction for specific cases[29]. Also in this case, wavelet transform's analysis can provide time-domain information, complementing CNNs in capturing temporal dependencies and enhancing audio classification.

4.5. Phase information

Spectrograms, as commonly used in audio processing, typically discard phase information and retain only the magnitude information. While this simplification enhances feature clarity and facilitates certain analysis tasks, it may result in the loss of valuable phase-related details, which can be essential for specific audio

processing applications. Thus, in scenarios where phase information plays a crucial role, such as sound source localization, phase-sensitive audio effects, or high-quality audio synthesis, the omission of phase data in spectrograms can limit the model's ability to accurately reconstruct and manipulate the audio signal.

To address this limitation, alternative representations that incorporate phase information can be explored, such as complex spectrograms or short-time Fourier transform (STFT)[30] with phase preservation. Additionally, advanced algorithms like the Griffin-Lim algorithm[31] can be utilized to reconstruct the phase from magnitude spectrograms when necessary, providing a more complete representation for certain audio processing tasks.

5. Practical applications

In this section, some practical applications of audio deep learning based on spectrograms are provided, based on the state of the art and selected to gain high accuracy and F1. For each study, we will describe the authors' perspective on a specific problem and describe their solution using spectrograms.

5.1. Randomized learning-based classification of sound quality using spectrogram image and time-series data

In the article titled "Randomized learning-based classification of sound quality using spectrogram image and time-series data: A practical perspective[32]", the authors discuss the use of randomized learning techniques to classify sound quality using spectrogram images and time-series data. The authors propose a method that involves obtaining data recorded from vehicle indoor noise and classifying them into three types of datasets through three preprocessing processes: spectrogram, variable-length time-series data, and up-sampling and interpolation scheme (USIS) data. To classify the sound quality of each dataset, they used CNN and LSTM networks for deep learning.

5.2. Multi-channel spectrograms for speech processing applications

In a research paper titled "Multi-channel spectrograms for speech processing applications using deep learning methods[33]", the authors propose a methodology to combine three different time/frequency representations of speech signals by computing multi-channel spectrograms continuous wavelet transform, Mel-spectrograms, and gammatone spectrograms and combining them into 3D-channel spectrograms to analyze speech in two different applications: automatic detection of speech deficits in cochlear implant users, and phoneme class recognition to extract phone-attribute features. To this aim, two different deep learning-based models are considered: convolutional neural networks and recurrent neural networks with convolution layers.

5.3. Applying image neural style transfer networks to audio spectrograms

In the research paper[34] titled "Sound transformation: Applying image neural style transfer networks to audio spectrograms", the authors purpose is to investigate whether audio spectrogram inputs can be used with image neural transfer networks to produce new sounds. Using musical instrument sounds as source sounds, the authors apply and compare three existing image neural style transfer networks for the task of sound mixing.

5.4. Environmental sound classification

In an article titled "Environmental sound classification using temporal-frequency attention based convolutional neural network[12]", the authors discuss the use of temporal-frequency attention-based convolutional neural network model (TFCNN) to classify environmental sounds. A method is proposed that involves obtaining data recorded from environmental sounds and classifying them into three types of datasets through three preprocessing processes: spectrogram, variable-length time-series data, and up-sampling and interpolation scheme (USIS) data. To classify the sound quality of each dataset, authors used CNN and LSTM networks for deep learning.

5.5. Speech-based emotion classification

In the paper titled "Deep learning techniques for speech emotion recognition: A review[35]", the authors present an introduction to various deep learning techniques with the aim of capturing and classifying emotional states from speech utterances. The authors tested the emotion-capturing capability of architectures such as convolutional neural networks (CNN) and long short-term memory (LSTM) using various standard speech representations such as mel spectrogram, magnitude spectrogram, and mel-frequency cepstral coefficients (MFCCs) on two popular datasets EMO-DB and IEMOCAP. The authors present experimental findings along with reasoning as to which architecture and feature combination is better suited for the purpose of speech emotion recognition.

5.6. Crowd-based emotional classification

In the paper "Emotional sounds of crowds: Spectrogram-based analysis using deep learning[5]", the authors propose a technique based on the generation of sound spectrograms from fragments of fixed length, extracted from original audio clips recorded in high-attendance events, where the crowd acts as a collective individual. Transfer learning techniques are used on a convolutional neural network, pre-trained on low-level features using the well-known ImageNet extensive dataset of visual knowledge. The original sound clips are filtered and normalized in amplitude for a correct spectrogram generation, on which they fine-tune the domain-specific features. Experiments held on the finally trained convolutional neural network show promising performances of the proposed model to classify the emotions of the crowd and compare results on different frequency/amplitude techniques (i.e., Mel, Bark, Log, Erb).

5.7. Edge emotion recognition

In a research paper titled "Edge emotion recognition: Applying fast Fourier transform on speech Mel spectrograms to classify emotion on a Raspberry Pi for near real-time analytics[36]", the authors examine audio files from five important emotional speech databases and visualize them in situ with dB-scaled Mel spectrograms using TensorFlow and Matplotlib. Audio files are transformed using a fast Fourier transform and fed into a convolutional neural network for classification.

5.8. Automated emotion recognition for groups

The article titled "Automatic emotion recognition for groups: A review[37]" aims to summarize and describe research on the topic of automatic group emotion recognition. In recent years, the topic of emotion analysis of groups or crowds has gained interest, with studies performing emotion detection in different contexts, using different datasets and modalities (such as images, video, audio, and social media messages), and taking different approaches. The authors suggest that research should test on multiple, common datasets and report on multiple metrics, when possible, to ensure clear, replicable, and comparative studies. They also suggest that an area of interest for future work is to build systems with more real-world application possibilities while having higher robustness and working with datasets with reduced biases. The previous work on crowd-sound emotions is cited in this review as the only reliable study on emotions from the sound of the crowd: this work uses spectrogram CNN-based classification.

6. Conclusion

We present a method for sound categorization that makes use of image-processing methods applied to audio spectrogram representations. The research shows how to improve feature extraction and categorization performance by integrating known image processing methods such as filtering, segmentation, and pattern recognition. Convolutional neural networks (CNNs) are used in the suggested approach to extract and categorize time/frequency features from spectrograms, taking advantage of their hierarchical feature learning and resilience to translation and scale changes, and of the possibility to apply transfer learning. The conversion

of auditory data to spectrograms has several advantages, including human interpretability, interoperability with image processing methods, and time/frequency resolution freedom. This method crosses the divide between image and audio processing. Through spectrograms, the article examines many similarities and various uses of shared methods between image processing and audio processing.

Acknowledgments

The author thanks both the anonymous external reviewer and the anonymous reviewer from the editorial board of the journal for the valuable insights that allowed to add so many interesting details to the paper.

VF is supported for this work by the University of Perugia, Department of Mathematics and Computer Science, and the KitLab research group.

Conflict of interest

The author declares no conflict of interest.

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 2017; 60(6): 84–90. doi: 10.1145/3065386
2. Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition. *arXiv* 2014; arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556
3. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27–30 June 2016; Las Vegas, NV, USA. pp. 770–778.
4. Thiery S, Nyiri E, Gibaru O, Boots B. Combining pretrained CNN feature extractors to enhance clustering of complex natural images. *Neurocomputing* 2021; 423: 551–571.
5. Franzoni V, Biondi G, Milani A. Emotional sounds of crowds: Spectrogram-based analysis using deep learning. *Multimedia Tools and Applications* 2020; 79: 36063–36075. doi: 10.1007/s11042-020-09428-x
6. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 18 August 2009; Miami, FL, USA. pp. 248–255.
7. Szegedy C, Ioff S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI conference on artificial intelligence; 4–9 February 2017; San Francisco, California USA. pp. 4728–4284.
8. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE conference on computer vision and pattern recognition; 7–12 June 2015; Boston, MA, USA. pp. 1–9.
9. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. In: Proceedings of the IEEE; November 1998. pp. 2278–2324.
10. Huang J, Chen B, Yao B, He W. ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE Access* 2019; 7: 92871–92880. doi: 10.1109/ACCESS.2019.2928017
11. McAllister T, Gambck B. Music style transfer using Constant-Q transform spectrograms. In: Martins T, Rodríguez-Fernández N, Rebelo SM (editors). *Artificial Intelligence in Music, Sound, Art and Design*, Proceedings of the 11th International Conference, EvoMUSART 2022; 20–22 April 2022; Madrid, Spain. Springer International Publishing; pp. 195–211.
12. Mu W, Yin B, Huang X, et al. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports* 2021; 11(1): 21552. doi: 10.1038/s41598-021-01045-4
13. Lin J, Li J, Duan G, Ming J. Wavelet features and gaussian process classifiers for face recognition. *Information Computing and Automation* 2008; pp. 47–50. doi: 10.1142/9789812799524_0013
14. Pasternack RM, Qian Z, Zheng JY, et al. Highly sensitive size discrimination of sub-micron objects using optical Fourier processing based on two-dimensional Gabor filters. *Optics Express*2009; 17(14): 12001–12012. doi: 10.1364/oe.17.012001
15. Jolliffe I. Principal component analysis. *Encyclopedia of Statistics in Behavioral Science* 2005. doi: 10.1002/0470013192.bsa501
16. Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Networks* 2000; 13(4–5): 411–430. doi: 10.1016/S0893-6080(00)00026-5
17. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS'00); 1 January 2000; Cambridge, MA, USA. pp. 535–541.
18. Weickert J. *Anisotropic Diffusion in Image Processing.* B. G. Teubner (Stuttgart); 1998.

19. Kelemenov T, Benedik O, Koláriková T, et al. Signal noise reduction and filtering. *Acta Mechatronica–International Sciencefic Journal about Mechatronica* 2020; 5(2): 29–34. doi: 10.22306/am.v5i2.65
20. Nguyen D, Widrow B. The truck backer-upper: An example of self-learning in neural networks. In: *Advanced Neural Computers*. NorthHolland; 1990. pp. 11–19.
21. Xiong F, Zhang Z, Ling Y, Zhang J. Image thresholding segmentation based on weighted Parzen-window and linear programming techniques. *Scientific Reports* 2022; 12(1): 13635. doi: 10.1038/s41598-022-17818-4
22. Cheng Z, Wang J. Improved region growing method for image segmentation of three-phase materials. *Powder Technology* 2020; 368: 80–89. doi: 10.1016/j.powtec.2020.04.032
23. Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. *International Journal of Computer Vision* 2004; 59: 167–181. doi: 10.1023/B:VISI.0000022288.19776.77
24. Wyse L. Audio spectrogram representations for processing with convolutional neural networks. In: Proceedings of the First International Workshop on Deep Learning and Music Joint with IJCNN; 17–18 May 2017; Anchorage, US. pp. 37–41.
25. Pal M, Roy R, Basu J, Bepari MS. Bind source separation: A review and analysis. In: Proceedings of the 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE); 25–27 November 2013; Gurgaon, India. pp. 1–5.
26. Zhang D. 2019. Wavelet transform. *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval*. Springer Cham; 2019. pp. 35–44.
27. Ding Y, Jia M, Miao Q, Cao Y. A novel time—Frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing* 2022; 168: 108616. doi: 10.1016/j.ymssp.2021.108616
28. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation* 2019; 31(7): 1235–1270. doi: 10.1162/neco_a_01199
29. Adavanne S, Politis A, Virtanen T. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In: Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO); 3–7 September 2018; Rome, Italy. pp. 1462–1466.
30. Parchami M, Zhu WP, Xhampagne B, Plourde E. Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits and Systems Magazine* 2016; 16(3): 45–77. doi: 10.1109/MCAS.2016.2583681
31. Perraudin N, Balazs P, Søndergaard PL. A fast Griffin-Lim algorithm. In: Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics; 20–23 October 2013; New Paltz, NY, USA. pp. 1–4.
32. Kang Y, Lee J. Randomized learning-based classification of sound quality using spectrogram image and time-series data: A practical perspective. *Engineering Applications of Artificial Intelligence* 2023; 120: 105867. doi: 10.1016/j.engappai.2023.105867
33. Arias-Vergara T, Klumpp P, Vasquez-Correa JC, et al. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications* 2021; 24: 423–431. doi: 10.1007/s10044-020-00921-5
34. Liu X, Delany SJ, McKeever S. Sound transformation: Applying image neural style transfer networks to audio spectrograms. In: Vento M, Percannella G (editors). *Computer Analysis of Images and Patterns*. Springer; 2019.
35. Pandey SK, Shekhawat HS, Prasanna SM. Deep learning techniques for speech emotion recognition: A review. In: Proceedings of the 2019 29th International Conference Radio elektronika (RADIOELEKTRONIKA); 16–18 April 2019; Pardubice, Czech Republic. pp. 1–6.
36. de Andrade DE, Buchkremer R. Edge emotion recognition: Applying fast Fourier transform on speech Mel spectrograms to classify emotion on a Raspberry Pi for near real-time analytics. doi: 10.21203/rs.3.rs-2198948/v1
37. Veltmeijer EA, Gerritsen C, Hindriks KV. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing* 2023; 14(1): 89–107. doi: 10.1109/TAFFC.2021.3065726