EXPLORING SEAM CARVING AS A DATA AUGMENTATION TECHNIQUE FOR ACOUSTIC CLASSIFICATION TASKS

Alicja Misiuda

University of Washington – AGH University of Krakow

ABSTRACT

A lack of quality data often can inhibit machine learning training. Datasets that are small and not diverse lead to undesirable results, all which slow down the task at hand. To improve training result, the technique of data augmentation has gained attention for its ability to enhance data at hand without needing additional external data. Existing data augmentation techniques have been proven to retain positive results, but a lack of variety introduces the potential for new strategies. The seam carving algorithm for image resizing can be modified to fit in an audio context and tested for effectiveness as a data augmentation technique for audio classification tasks. The algorithm was evaluated on the DCASE 2020 Task 1a data set with the goal of surpassing the ResNet accuracy of 74.6% found from a previous paper that attempted the original DCASE 2020 Task 1a challenge [1].

Index Terms— Data augmentation, seam carving, acoustic scene classification, resnet, dynamic programming

1. INTRODUCTION

Data augmentation has become a critical technique in machine learning to combat the training difficulties that arise from small datasets. CNNs are some of most common models used for audio classification and the quality of its information retention reduces with these small datasets. Now, there already exist various types of data augmentation strategies as outlined in [2], where they enhance the diversity of training datasets without requiring additional external data. Specifically, pitch shifting has been found to be beneficial for ship audio classification [3], along with spectrogram augmentation for animal audio classification [4]. These techniques alter the representation of the audio signal in ways that introduce variations that aim to prevent the models from memorizing training data and losing performance on new data.

Seam carving was originally introduced in 2007 as a content-aware image resizing technique [5]. The seam carving algorithm intelligently resizes images by removing seams—paths of least importance based on pixel content—thereby preserving essential structures while discarding unnecessary information. This technique operates on the 2D array representation of an image, allowing for understandable methodology and representation for repeatability. Additionally, the algorithm possesses the ability to preserve the most important image content. With the possibility of the 2d array representation, this algorithm could be altered to be applicable for audio, and perhaps an intriguing alternative to traditional audio augmentation methods.

The spectrogram, a 2D representation of audio data where one axis corresponds to time and the other to frequency, creates the grounds for the application of seam carving within the audio domain. By treating the spectrogram as an image, the seam carving technique can manipulate audio signals by identifying and removing less critical time-frequency components. This raises an important question: can seam carving serve as a viable data augmentation technique for audio classification tasks? Specifically, how does it compare to existing methods like pitch shifting and time stretching in terms of improving acoustic classification performance?

The objective of this study is to investigate how viable the seam carving algorithm is as a data augmentation method for acoustic scene classification. The aim to determine whether it can compete with or even outperform established techniques like pitch shifting and time stretching. To this end, I applied the seam carving algorithm to the DCASE 2020 Challenge Task 1a dataset, a benchmark dataset for acoustic scene classification.

Although seam carving has previously been utilized in audio processing for tasks such as extracting sinusoidal components from sound spectrograms [6] and creating seamless listening experiences in music [7], its application as a data augmentation technique for audio classification remains unexplored. This paper seeks to fill that gap by demonstrating the potential of seam carving to enhance audio classification models, providing a novel approach to data augmentation in this domain.

2. RELATED WORK

Previous work involved utilizing pitch shifting and spectrogram augmentation for different audio classification applications. The two papers concerned with this type of data augmentation were exploring how their techniques would improve audio classification accuracy for their respective datasets. The proposed methods stemmed from already existing data augmentation techniques and were focused on testing the new combinations.

In recent years, deep learning techniques represented by CNNs have been utilized for ship-radiated noise classification which deals in the realm of underwater acoustics and marine research. The improved pitch shifting method enhances ship-radiated noise classification by acting as a robust data augmentation technique [3]. Traditional pitch shifting methods often fail to introduce sufficient variability in training data, leading to limited improvements in model performance. IPS addresses this by dynamically varying the pitch over time, which better simulates real-world variations in ship noise, such as changes in speed or operational state. This approach significantly increases the diversity of training samples, helping to prevent overfitting and improving model generalization. As a result, IPS, especially when combined with time stretching, boosts classification accuracy and F1 scores across datasets like DeepShip and ShipsEar, making it a valuable tool for improving ship noise classification performance. Additionally, a study on data augmentation techniques for animal audio classification discovered spectrogram augmentation as an effective data augmentation technique used in animal sound classification by modifying the visual representations of audio signals [4]. This approach involves applying various transformations, including pitch shifts, time shifts, and adding noise, directly to the spectrogram images rather than the raw audio. These manipulations create a wider variety of training samples that mimic the natural variability found in real-world animal sounds. As a result, the models trained on these augmented spectrograms are more robust, better generalize to new data, and achieve improved classification performance across different animal sound datasets.

3. METHODOLOGY

To achieve content-aware resizing of images, I employed a dynamic programming approach to identify and manipulate vertical seams of minimal energy within the image. A seam is defined as a connected path of pixels that spans from the top to the bottom of the image, where each pixel in the seam is directly adjacent to the one above it (i.e., the pixels are vertically or diagonally connected). This method focuses on calculating and identifying the optimal vertical seam with the least cumulative energy, which minimizes the visual impact of resizing operations.

3.1 Initialization and Energy Calculation

Begin by computing the energy of each pixel in the image, stored in an energy matrix. This energy matrix reflects the importance of each pixel, with lower values indicating less critical content. Then initialize two matrices: seam_energies and back_pointers. The seam_energies matrix is used to store the cumulative minimum seam energies for each pixel, while the back_pointers matrix tracks the optimal paths by recording the indices of the contributing pixels from the previous row.

3.2 Dynamic Programming Seam Extraction

The first row of seam_energies (represented s in equation) is directly set to the corresponding values in energies, as there are no prior rows influencing these values. For each subsequent row, the algorithm iteratively computes the cumulative seam energy for each pixel based on the minimum seam energies of the three possible preceding pixels from the row above: left diagonal, vertical, and right diagonal. The cumulative energy of a pixel is computed as:

$$\mathrm{s[y,x]} = \mathit{energies}[\mathrm{y,x}] + \min\,(s[y-1, x-1, s)$$

The corresponding back_pointers entry for each pixel is updated to reflect the index of the pixel from the previous row that contributed to this minimum value, thereby storing the optimal path incrementally.

Upon completion of the seam energy calculations for all rows, the minimum value in the last row of seam_energies indicate the end of the optimal vertical seam. The seam itself is reconstructed by backtracking from this minimum value using the back_pointers matrix, tracing the path of minimal energy back to the top of the image.

3.3 Application and Efficiency

In my approach to seam carving, I utilized dynamic programming to efficiently identify and manipulate seams with minimal energy. To benchmark the effectiveness of this method, I compared it with a previous alternative approach using Dijkstra's algorithm, which is commonly employed for finding shortest paths in graphs.

The method computes the cumulative minimum seam energies in a bottom-up manner, storing these values in a matrix that represents the entire image grid. This approach benefits from the dynamic programming implementation as the time complexity results in O (height x width), where each pixel's energy computation is constant time, resulting in linear scalability relative to the image dimensions. The approach directly computes the optimal seam and leverages back pointers to reconstruct the seam path without additional overhead. Additionally, the method only requires matrices proportional to the size of the audio spectrogram, ensuring that memory usage remains manageable even for larger audio clips.

Dijkstra's algorithm, while versatile and applicable to various shortest path problems, involves a more generalized approach that is less efficient for grid-based problems like seam carving. It constructs a graph where each pixel is a node, and edges represent possible paths between adjacent pixels with weights corresponding to pixel energies. This approach has several limitations, as it operates with a time complexity of O ((V+E) log V), where V represent the vertices and E represent the edges. The need to construct and manage a graph adds significant overhead, including the complexities of handling data structures like priority queues for pathfinding. Unlike DP, which directly leverages the structured grid nature of the image, Dijkstra's algorithm does not inherently optimize for this structure, leading to redundant computations.

4. EXPERIMENTS

After the successful completion of the adapted algorithm, the experiments were conducted using the 10-stage ResNet structure found in a paper specific for the DCASE 2020 Task 1a [1] along with the original data set from the DCASE 2020 Challenge [8]. The testing was conducted with a set of procedures. The acoustic scenes audio dataset would be "seam carved" with a varying seam amount. The original audio clips have a size of (1025, 862) meaning that there were 1025 possible seams to be carved in the time domain and 862 possible seams to be carved in the frequency domain. Experiments were tested with 400 seams carved, a random number of seams carved for each audio clip between 300 and 900 seams, and a random amount of seams carved for each audio clip between 500 and 900 seams, all three experiments in the frequency domain. Furthermore, to compare the seam carving results to other data augmentation methods, another set of data was created using librosa's pitch shift function, setting the number of steps to be between -6 and 6. All these augmented audio files were tested individually as well as adding it to the original data.

5. RESULTS

The experiments yielded a set of validation accuracy metrics for each data augmentation method applied to the ResNet model structure, as depicted in Figure X. The baseline model using the original ResNet achieved the highest validation accuracy of 0.711, demonstrating its effectiveness compared to all tested data augmentation strategies.

Among the seam carving experiments, the augmentation with 400 seams carved in the frequency domain reached a validation accuracy of 0.682. This suggests a moderate level of enhancement but still trailing the original ResNet. The results were slightly lower for the mixed seam carving approaches: the experiment with a variable seam number between 300 and 900 (denoted as "mixed_sc") reached a validation accuracy of 0.656, while the approach with seams carved randomly between 500 and 900 seams (denoted as "mixed_sc_two") achieved a validation accuracy of 0.674.

Additionally, the pitch shifting data augmentation method, which involved altering the audio pitch between -6 and 6 steps, achieved a validation accuracy of 0.638. This value, while above the baseline of random seam carving ("rerun frequency 400" at 0.162), was still significantly lower than both the original ResNet and the seam carving approaches, suggesting that pitch shifting along may be less less effective for this specific task.

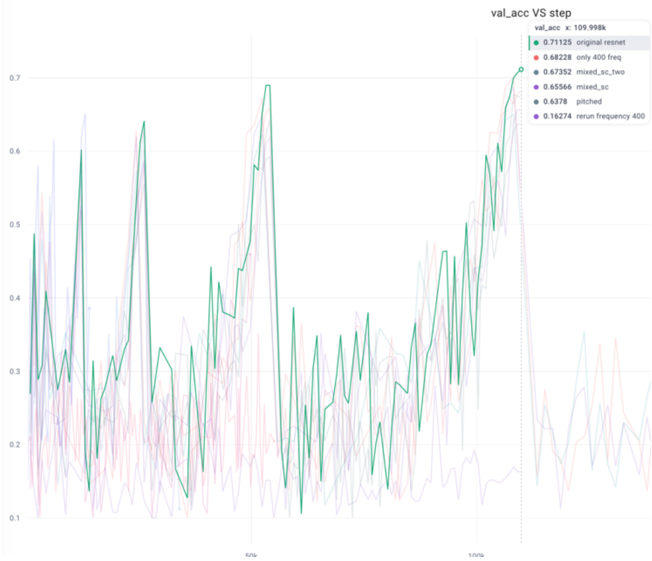

Figure 1: Validation Accuracy for Experiments

Overall, the experiments indicate that while data augmentation methods like seam carving and pitch shifting offer some improvements, none of them surpass the performance of the original ResNet on the DCASE 2020 Task 1a dataset. Future work may explore combining these augmentation techniques or tuning them further to enhance model performance.

6. CONCLUSIONS

The experimental results indicate that seam carving, as implemented solely in the frequency domain, does not significantly improve the performance of the ResNet model for the DCASE 2020 Task 1a dataset. The validation accuracy for seam carving methods, while somewhat effective, did not surpass that of the original ResNet model without augmentation. This suggests that the current approach to seam carving might have limited influence when applied independently.

Importantly, the scope of seam carving in this study was confined to the frequency domain. Exploring seam carving in the time domain could potentially yield different results, but this would necessitate modifications to the original ResNet structure to accommodate changes in the temporal parameters of the input data. Such adjustments remain a subject for future research.

Additionally, the potential of seam carving could be more fully realized when used in conjunction with other data augmentation techniques. Previous work has shown that spectrogram augmentation combining methods like time shift, pitch shift, and image rotation can effectively enhance model performance. Seam carving, therefore, could serve as an additional building block in a multi-faceted augmentation strategy, potentially contributing to improved robustness and generalization of audio classification models.

Future work will focus on expanding the application of seam carving to the time domain and integrating it with other augmentation methods to explore its full potential within a comprehensive data augmentation pipeline.

7. RELATION TO PRIOR WORK

Previous studies have demonstrated the use of data augmentation techniques, such as pitch shifting and spectrogram augmentation, to enhance audio classification accuracy across various datasets [3][4]. These methods primarily aimed to increase the diversity of training data, thereby improving model robustness and generalization. The pitch shifting methods applied in prior research often dealt with static alterations in audio features, which provided limited variability in training data and modest improvements in model performance. Notably, the Improved Pitch Shifting method advanced this concept by dynamically varying pitch over time, thus better mimicking real-world conditions, such as changes in ship speed or operational states in underwater acoustics applications.

While the existing literature focuses extensively on enhancing audio classification through augmentation techniques, my work introduces a novel approach by employing content-aware resizing of images using dynamic programming to identify and manipulate vertical seams of minimal energy. Unlike static data augmentation, which alters features indiscriminately across entire datasets, my approach selectively modifies the audio based on its content, preserving essential features while resizing. This is achieved through the seam carving algorithm, which calculates the optimal vertical seam with the least cumulative energy, effectively resizing spectrograms with minimal audio distortion.

In contrast to the approaches in audio classification, where augmentation techniques are applied uniformly across the dataset, the seam carving method adapts dynamically to the specific content of each spectrogram image This dynamic adaptation is analogous to the IPS method’s dynamic pitch variation, yet it extends beyond augmentation into the realm of content-aware resizing, a distinct application that addresses the preservation of signal integrity during resizing tasks. Furthermore, by leveraging the seam carving technique, which relies on dynamic programming principles to iteratively calculate and trace minimal energy paths.

Overall, my work diverges from previous methodologies by shifting the focus from audio to image processing while maintaining the underlying theme of content preservation and enhancement. This not only broadens the application of dynamic programming techniques to new domains but also establishes a unique pathway for future research into content-aware manipulation of digital media, reinforcing the interdisciplinary potential of these approaches.

8. REFERENCES

[1] H. Hu *et al.*, “A Two-Stage Approach to Device-Robust Acoustic Scene Classification.” Accessed: Sep. 09, 2024. Available: https://arxiv.org/pdf/2011.01447

[2] L. Ferreira-Paiva, E. Alfaro-Espinoza, V. Almeida, L. Felix, V. Rodolpho, and Neves, “A Survey of Data Augmentation for Audio Classification.” Available: https://www.sba.org.br/cba2022/wp-content/uploads/artigos_cba2022/paper_5085.pdf

[3] Xu Yuanchao, Cai Zhiming, and Kong Xiaopeng, “Improved pitch shifting data augmentation for ship-radiated noise classification,” *Applied Acoustics*, vol. 211, pp. 109468–109468, Aug. 2023, doi: https://doi.org/10.1016/j.apacoust.2023.109468.

[4] L. Nanni, G. Maguolo, and M. Paci, “Data augmentation approaches for improving animal audio classification,” *Ecological Informatics*, vol. 57, p. 101084, May 2020, doi: https://doi.org/10.1016/j.ecoinf.2020.101084.

[5] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 10, Jul. 2007.

[4] G. Capizzi, D. Rocchesso, and S. Baldan, “Streams as Seams: Carving trajectories out of the time-frequency matrix,” 2020. Accessed: Aug. 25, 2024. [Online]. Available: https://air.unimi.it/retrieve/f2113b8c-7013-401f-b1d5-8897ae5d1014/SMCCIM_2020_paper_60.pdf

[6] M. Covell and S. Baluja, “Seamless Audio Melding: Using Seam Carving with Music Playlists.” Accessed: Aug. 25, 2024. Available: http://www.esprockets.com/papers/mmedia.pdf

[7] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic Scene Classification in DCASE 2020 Challenge: Generalization Across Devices and Low Complexity Solutions,” in DCASE2020, 2020.

[8] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. *Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions.* In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 56–60. 2020. URL: https://arxiv.org/abs/2005.14623.