A report for AI project

"Data Augmentation for Text, Image, Sensor, and Audio Corpora"

Sristi Bhadani
Soham Joita

Supervised by
Prof. Dr.-Ing. Christian Bergler

Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

July 1, 2024

Abstract

Data augmentation is an effective strategy for increasing the robustness and generalizability of machine learning and deep learning models. It has a considerable impact in domains where getting large, diverse, and labeled datasets is challenging. The proposed study focuses on the development of data augmentation strategies for four different data types: text, sensor, image, and audio. Each modality poses various obstacles and ways for augmentation, therefore tactics must be tailored to the data's individual qualities. The primary goal of this research is to investigate various augmentation strategies for text, sensor, image, and audio data, in order to uncover methods that improve model performance, particularly when training data is low. This study also presents a thorough methodology for implementing augmentation tactics across various data types. For text data, approaches such as synonym replacement, random insertion, and back-translation were investigated. Sensor data augmentation techniques included jittering, scaling, and permutation. Rotation, flipping, color modifications, and blurring were some of the ways used to enrich image data. Audio data was processed using techniques such as time stretching, pitch shifting, and noise adding. To attain the best results, augmentation techniques must be tailored to the individual qualities of text, sensor, image, and audio. These findings are a great resource for practitioners looking to implement data augmentation in machine learning, emphasizing the potential for performance improvements. Overall, this paper presents a complete examination of data augmentation strategies, as well as insights into their application and impact across various data modalities.

1 Introduction and Motivation

In artificial intelligence and deep learning, data quality and variety are critical to model success across domains. Image, audio, text, and sensor data are crucial and have numerous uses in machine learning and deep learning. Acquiring labeled data in sufficient quantity and diversity is difficult, sometimes resulting in models that overfit or do not generalize effectively. Data augmentation is a machine learning technique that artificially increases the size and diversity of a training dataset without gathering any new data. The approach aids in the creation of modified versions of existing data points, which improves model training. To generate new textual variations, methods such as synonym replacement and back-translation are used. Sensor data augmentation uses techniques like jittering and scaling to increase the variety of sensor readings. Image data augmentation involves changes such as rotation and color modifications, whereas audio data augmentation employs noise addition and pitch shifting.

Despite the efficiency of existing augmentation strategies, there is a significant need in the literature for a unified framework that can combine augmentation techniques across different data modalities. Existing research frequently focuses on single data categories, leaving a gap in our understanding of how different augmentation procedures operate when applied to text, sensor, image, and audio datasets.

The primary goals of this research are to create a unified augmentation library that can improve audio, text, image, and sensor corpora by incorporating advanced augmentation techniques, as well as to design a scalable and efficient framework for generating augmented data samples while maintaining semantic coherence and integrity.

The proposed solution enhances the research field of machine learning, especially in areas where data is limited or expensive to gather. This study contributes to the creation of more robust and generalizable models. Our system, which provides a consistent method for augmenting multiple data modalities, can enable future research in data augmentation, transfer learning, multi-modal training, and other areas.

The paper is organized as follows: Section 3 describes the data and its sources. Section 4 gives a thorough analysis of existing augmentation strategies for picture, audio, text, and sensor data domains. Section 5 covers how we designed and implemented our unified data augmentation architecture. In Section 6, we assess the framework's performance on a range of machine learning tasks. Finally, Section 7 analyzes the consequences of our findings and suggests areas for future investigation.

2 Related Work

Recent advances in several disciplines of machine learning have fueled creative techniques to data augmentation, considerably improving model performance and reliability. Sennrich et al. [6](2016) presented back translation strategies in Natural Language Processing (NLP) to improve neural machine translation using monolingual data. Wei and Zou [10] (2019) proposed simple but effective methods for text classification tasks, such as synonym replacement and random perturbations. Devlin et al. [3] (2019) demonstrated BERT's ability to generate high-quality augmented text, enriching training datasets. Junczys-Dowmunt et al. [4] (2018) used machine translation to correct grammatical errors, highlighting the range of augmentation tactics in NLP.

Wang et al.'s [9] research article "Sensor Data Augmentation by Resampling for Contrastive Learning for Human Activity Recognition" discusses numerous sensor data aug-

mentation approaches. These techniques include jittering, scaling, magnifying, rotating, inverting, reversing, permutation, temporal warping, cropping, and shuffling. They also provide a novel resampling method that combines upsampling and downsampling to imitate changing sensor frequencies and improve augmentation efficacy. These strategies improve the reliability of contrastive learning frameworks for human activity recognition problems.

Meanwhile, in computer vision, picture data augmentation is critical for improving deep learning model performance, particularly when labeled data is limited. Aside from classic geometric changes and color space tweaks, recent improvements have included the incorporation of Generative Adversarial Networks (GANs) and neural style transfer algorithms. These methods provide very realistic augmented images, which considerably improve model generalization in applications such as medical imaging, where large datasets are sometimes difficult to collect. Connor Shorten and Khoshgoftaar's [2] survey ("A Survey on Image Data Augmentation for Deep Learning") highlights the transformative impact of these techniques on computer vision tasks and their growing significance in advancing deep learning applications.

Furthermore, in recent years, audio augmentation has emerged as an important strategy for improving the performance of deep learning models in a variety of audio processing applications, such as voice recognition and acoustic scene classification. Comprehensive reviews have established the effectiveness of techniques such as noise addition, pitch shifting, time stretching, and vocal tract length perturbation. Furthermore, time-domain augmentation techniques such as mixing and cropping have demonstrated efficacy in auditory scene classification tasks. These augmentation procedures are critical for developing diverse and representative training datasets, resulting in more robust and generalizable models. Ravi et al. [5] made significant contributions in this subject with their work and reviews on comprehensive data augmentation strategies for speech and audio processing.

These broad advancements highlight the changing environment of data augmentation techniques, which are driving gains in model resilience, generalization capabilities, and performance across multiple fields of machine learning and AI. By extending existing approach, we hope to contribute to the improvement of multimodal data augmentation techniques and the construction of more robust models.

3 Data Materials

This project requires four different types of datasets. The datasets are sourced from the Kaggle platform and zedge.net.

3.1 Text dataset

The name of this dataset is "Reviews,[8]" and it is in CSV format. This dataset contains approximately 500,000 reviews of Amazon fine foods. It includes product and user information, ratings, and plain text reviews. Additionally, it encompasses reviews from all other Amazon categories. We are considering Summary column from the dataset for our research.

3.2 Sensor dataset

For this project, we use live smoke sensor dataset [7], which is in CSV format. This data is collected from IoT devices deployed in various environments such as normal indoor, normal outdoor, indoor wood fire, firefighter training areas, indoor gas fire, outdoor wood, coal, and gas grill, and outdoor high humidity areas. The dataset contains 60,000 readings with UTC timestamps.

3.3 Image dataset

We conducted experiments using three different datasets to get variety of image data. Hence, it comprises of animal, human being, vegetable data which were visualised.

3.4 Audio dataset

We have used the ringtone [1] for this dataset, took sample of 3 audios to augment it.

4 Methodology

The methodology of the project relies around the creation of a unified data augmentation framework for image, audio, text, and sensor corpora. The technique enables the design and implementation of augmentation algorithms specific to each data modality, guaranteeing that the augmented data introduces unpredictability and simulates real-world settings. In our work, we used a variety of text augmentation approaches to enrich and diversity our dataset, which improved the robustness and generalizability of our models. Key strategies were synonym replacement using WordNet, which increased lexical variety while keeping semantic meaning, noise addition to simulate typographical errors, and training models to manage modest text perturbations. Methods such as random insertion, deletion, and swapping increased text variety. In addition, back translation and style transfer procedures, as well as advanced generation methods based on probabilistic and graph-based models, produced various and contextually coherent text variations. The BERT augmentation approach and grammatical correction guaranteed contextual integrity and text accuracy, respectively.

For sensor data, a clipping mechanism ensures that values remain within the original range, preventing unrealistic results from noise addition. Gaussian noise is produced using a normal distribution with a mean of 0.0 and a standard deviation of 0.01, whereas uniform noise employs lower and upper boundaries. Jittering is achieved using the Gaussian distribution. The GenerateRandomCurves function creates smooth, random curves using cubic splines, resulting in realistic variations. The Permutation function randomly permutes data segments to increase diversity by dividing each row, swapping segments, and reassembling them in a different sequence.

The augmentation function dynamically performs a variety of transformations to picture data based on the configuration specified. The Compose technique successively combines numerous transformations to create complicated augmentation pipelines. Resize scales photos while keeping their aspect ratio. RandomHorizontalFlip and RandomVerticalFlip introduce orientation variety by flipping images at a predetermined probability. RandomRotation improves rotational robustness by rotating images to defined degrees. Grayscale reduces color channels by converting images to grayscale. ColorJitter modifies

brightness and hue to simulate various lighting conditions. RandomAffine transforms image geometry by applying transformations such as translation, scaling, and rotation. GaussianBlur simulates out-of-focus effects with a Gaussian kernel. RandomPosterize decreases color bit depth, resulting in varying intensity levels. CenterCrop crops an image's middle portion to highlight the main topic. RandomPerspective introduces perspective distortions, resulting in several views. RandAugment, AugMix, and TrivialAugmentWide are techniques for automating augmentation policies that use random and mixed modifications to improve image variability and resilience. These methods ensure a large range of transformations, making it easier to train machine learning models that generalize well across varied circumstances.

SSeveral audio augmentation techniques were used to generate diverse and representative datasets. The `apply_augmentation` function dynamically applies different approaches based on the specified configurations. The `add_noise` function generates random noise controlled by a noise factor to simulate various situations. The `change_speed` function modifies playback speed, resulting in tempo and rhythm alterations. The reverse audio technique reverses the audio signal and establishes a backward playback effect. The `slow_down_audio` function lowers the playback rate, making the audio sound more drawn out. The `add_echo` method adds an echo effect with delay and decay parameters, resulting in a reverberation effect. The `pitch_shift` function changes the pitch by a specified number of steps, affecting the apparent frequency. The `time_masking` approach masks a portion of the audio to simulate missing data and improve robustness against dropout. The `add_sonic_boom_effect` method adds a delayed, amplified signal with decay to simulate a sudden, strong sound impact. Each augmentation increases the diversity and durability of audio samples for machine learning applications, assuring realistic variances while maintaining data integrity.

5 Experiments

The tests are carried out to assess the effectiveness of generalizing the suggested data augmentation methodology across image, audio, and text data. The research focuses on enhanced data. The investigations were performed in a computational setting. Python is the programming language used in the software environment. Pytorch, numpy, pandas, transformers, matplotlib, sound file, librosa, sentencepiece, and other libraries are worked with. Furthermore, libraries for data processing and analysis include NumPy, pandas, and scikit-learn. Virtual environments are used to manage dependencies, ensuring reproducibility. The specific versions of software packages and setups are documented so that the experimental results can be replicated.

In the textual data augmentation project, we used a variety of strategies to improve our dataset's diversity and robustness. Synonym replacement, adding noise, random insertion, random deletion, random swap, grammar correction, BERT augmentation, text generation, back translation, paraphrasing, style transfer, text generation for graph structures, hierarchical text generation, conditional text generation, and stochastic text generation are among the techniques employed. The following table shows the outcomes of the input data column of "delight says it all" applying several augmentations, exhibiting the alterations done to the original text:

Techniques	Results
Synonym replacement	joy order IT altogether
Noise	delight bays Oe alp
Random Insertion	k delight says k it x all
Random Deletion	delight says it
Random Swap	says it delight all
Paraphrasing	enchant `sound_out` IT totally
`Style_Transfer_Nmt`	DELIGHT SAYS IT ALL
`Stochastic_Text_Generation`	it says it says
Masking	delight [MASK] [MASK] all
`Grammar_Correction`	delight says it all
`Text_Generation`	delight says it all
`Conditional_Text_Generation`	delight says it all
`Word_Shuffling`	it says all delight
`Bert_Augmentation`	delight [unused583] [unused341] all
`Hierarchical_Text_Generation`	IT pronounce IT completely
`Back_Translation`	delight says it all

Table 1: `Text_results`

To improve the stability and diversity of our sensor datasets, we applied a number of sensor data augmentation methodologies in our experimental scenario. These strategies included using Gaussian noise to replicate real-world variances and measurement mistakes, as well as introducing homogenous noise to cover a wide range of sensor data. We also used data augmentation (DA) jitter to randomly shift the sensor data points and DA scaling to change the amplitude of sensor readings, imitating various operational situations. In addition, we produced random curves to introduce synthetic but realistic changes and used DA permutation to shuffle parts of the data, guaranteeing that the model learns to recognize patterns regardless of order.Each of these augmentation strategies was used on our sensor datasets to produce a comprehensive and durable dataset for training our models. The outcomes of these augmentations have been presented in tabular form.

Index	Temperature[C]	Humidity[%]
1	20.0000	57.3600
2	20.0150	56.6700
3	20.0290	55.9600
4	20.0440	55.2800

Table 2: Original Data

Index	Temperature[C]	Humidity[%]
1	19.9959	57.3516
2	20.0143	56.6609
3	20.0200	55.9633
4	20.0570	55.2811

Table 3: Gaussian Noise

Index	Temperature[C]	Humidity[%]
1	19.9961	57.3662
2	20.0234	56.6743
3	20.0260	55.9570
4	20.0532	55.2787

Table 4: Uniform Noise

Index	Temperature[C]	Humidity[%]
1	20.0376	57.3768
2	19.9591	56.6627
3	20.0590	55.9196
4	19.9931	55.3009

Table 5: Jitter

Index	Temperature[C]	Humidity[%]
1	21.0203	53.9770
2	21.0361	53.3277
3	21.0508	52.6596
4	21.0666	52.0197

Table 6: Scaling

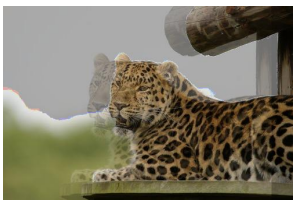
Index	Temperature[C]	Humidity[%]
1	21.1005	33.7349
2	21.1164	33.3288
3	21.1312	32.9109
4	21.1470	32.5107

Table 7: GenerateRandomCurves

Index	Temperature[C]	Humidity[%]
1	20.0804	57.3594
2	20.0527	56.7503
3	20.0085	56.0304
4	20.0132	55.2363

Table 8: Permutation

In our experimental setup for image data augmentation, we implemented a variety of approaches such as resize, horizontal flip, rotation, grayscale, vertical flip, color jitter, random affine transformations, Gaussian blur, random posterization, center cut, random perspective, and elastic transform. The outcomes of these augmentations, given in a flow manner, demonstrate the visual modifications made to the images:



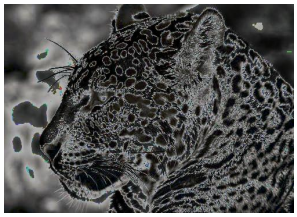
(a) Augmix

(b) R-Perspective

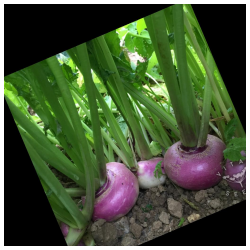
(c) Colorjitter

(d) Gaussianblur

(e) Grayscale



(f) R-Augment



(g) R-affine

(h) R-HorizontalFlip

(i) R-Posterize

(j) R-Rotation

(k) R-VerticalFlip

(l) TrivialAugWide

(m) CenterCrop

To improve the durability and diversity of our audio datasets, we used a number of audio data augmentation methodologies in our experimental setting. These strategies included introducing noise to imitate real-world circumstances with background sounds like traffic or people, reversing audio to test the model's ability to distinguish temporal elements, and varying the playback speed to accommodate for varied speaking speeds or musical tempos.In addition, we slowed down audio to highlight detailed details, added echo effects to simulate diverse acoustic surroundings, and used pitch shifting to handle fluctuations in vocal pitch or musical key. Each of these augmentation strategies was applied to our audio datasets, resulting in a complete and resilient dataset for training our models. The findings of these augmentations were processed and safely kept in Google Drive, ensuring that they are backed up and conveniently accessible for future research and sharing.

6 Results and Discussion

6.1 Text

In our study, we applied a number of text-augmentation techniques to increase dataset diversity and model resilience. Combination tactics add context but may cause redundancy. Synonym Replacement maintains meaning with other phrases, but it can be difficult. Adding noise generates typographical errors, which increases the reliability of real-world data. Random insertion and deletion increase diversity but frequently disrupt syntax and meaning. Random swaps add grammatical variety but degrade structure. Masking is an excellent method for predicting missing words, but it leaves the text incomplete. Grammar correction ensures readability and coherence. Generating and paraphrasing produce new sentences that may stray from their original context. Back translation preserves meaning with modest variations, whereas stochastic and hierarchical text production differ in quality. Text creation from graph structures gives logical flow, although it is complex. BERT Augmentation employs deep models to generate robust results, whereas Style Transfer NMT alters style features. Conditional text generation maintains relevance and context. Overall, Back Translation and Paraphrasing are the best at preserving meaning while allowing for diversity, however Grammar Correction is essential for readability and coherence.

6.2 Sensor

Gaussian noise, which adds normally distributed random noise, creates slight deviations while retaining overall trends, making it helpful for imitating sensor noise without drasti-

cally affecting the data distribution. See Table 3. Uniform noise, as described in Table 4, adds random values uniformly distributed within a particular range, producing minimum, controlled deviations akin to Gaussian noise. Table 5 describes Jitter, which provides minor, random fluctuations using a Gaussian distribution for multiplicative changes, while keeping general patterns. Scaling, seen in Table 6, multiplies data by a factor, resulting in considerable uniform increases. It is ideal for modeling broader environmental changes, but may distort the original distribution.The technique Generate Random Curves, as stated in Table 7, introduces complicated, non-linear changes, resulting in significant deviations that are beneficial for stress-testing models, but may overcomplicate the data. Permutation, as described in Table 8, shuffles data segments while maintaining integrity through controlled randomness. Each technique has advantages: Gaussian and uniform noise for minor perturbations, jitter for small fluctuations, scaling for larger changes, Generate Random Curves for non-linear variations, and permutation for controlled randomness. Among these, Gaussian noise is the most effective at retaining the original distribution while adding genuine variability.

6.3 Image

In this study, we used a variety of picture augmentation techniques to increase the resilience and diversity of our dataset. The Compose technique combines many transformations to generate sophisticated augmentation pipelines. We used Resize to standardize the image dimensions, assuring uniformity. Figures 1h and 1k demonstrate how RandomHorizontalFlip and RandomVerticalFlip introduced random flips with defined probabilities (p=1.0), introducing orientation variety. Figures 1j and 1g show the use of RandomRotation and RandomAffine, which give geometric distortions such as random rotations within 30 degrees and affine transformations with translation parameters of [0.1, 0.1] and scaling of [0.8, 1.2]. Grayscale (1e) is a transformation that reduces image complexity by removing color information. ColorJitter (1c) generated random changes in brightness (0.5) and hue (0.3) to simulate various lighting conditions. GaussianBlur (1d) creates a soft blur effect adjusted by kernel size ([5, 9]) and sigma ([0.1, 0.5]) to simulate out-of-focus conditions. RandomPosterize (1i) lowered color resolution and produced a posterization effect with 2 bits. CenterCrop (1m) focuses on the core section of the image while preserving important elements. RandomPerspective(1b) applied perspective distortions with a distortion scale of 0.6, whereas TrivialAugmentWide(1l) applies a single random transformation from a large set, such as rotations or color adjustments, to each image. RandAugment (1f) uses a predetermined set of augmentations performed in random order with changing intensity to enhance diversity without substantial adjustment. AugMix (1a) randomly combines numerous augmentations and blends them with the original image. These augmentation strategies helped to create a more generic and adaptable machine learning model that can perform well in a variety of unpredictable real-world scenarios.

6.4 Audio

IIn this study, we employed a variety of audio augmentation approaches to improve dataset resilience and variability. The function `mp3_to_wav` uses the pydub library to convert audio files from MP3 to WAV, standardizing the format for processing. The `add_noise` technique uses random noise (noise factor 0.3) to simulate real-world situations

with background noise. The `reverse_audio` approach reversed the audio signal, which helped the model acquire invariant audio properties. To mimic varied playback speeds, the `change_speed` method doubled the speed (factor 2.0), and the `slow_down_audio` function halved it (rate 0.5), allowing the model to recognize varying tempos. The `add_echo` function created an echo effect with a 0.5-second delay and a 0.6 decay factor to simulate reflective environments and increase data diversity. Furthermore, the `pitch_shift` approach shifted the pitch by two steps, affecting perceived frequency. The `time_masking` technique used a mask over 50% of the audio to simulate missing data and improve robustness against dropout. The `add_sonic_boom_effect` created a dramatic effect with a 0.05-second delay, an increase factor of 10, and a decay of 0.6 to simulate a loud sound impact. These strategies boosted the model's capacity to generalize to a wide range of real-world scenarios by infusing variety and complexity into the training data.

7 Summary, Conclusion, and Future Work

7.1 Summary

This work thoroughly examined and developed data augmentation techniques for four major data types: text, sensor, image, and audio. Text data diversification procedures such as synonym substitution, random insertion, and back-translation were found to be useful in enhancing model generalizability. Techniques including jittering, scaling, and permutation were utilized in sensor data to create unpredictability, resulting in more robust models. Image data augmentation including rotation, flipping, color changes, and blurring significantly improved model performance by providing a wide range of visual alterations. Time stretching, pitch shifting, and noise addition were successfully performed on audio data, increasing the robustness of audio processing models.

7.2 Conclusion

The findings of this study demonstrate the efficacy of tailored data augmentation strategies in boosting the performance and robustness of machine learning models across a wide range of data types. The methodical implementation and evaluation of these tactics shown their ability to increase model generalization, particularly in circumstances with little training data. The unified architecture built takes a comprehensive approach to data augmentation, addressing challenges such as data scarcity and unpredictability. Despite a number of limitations, including potential biases introduced by some procedures and computational resource constraints, the study stresses the importance of well-designed augmentation strategies for boosting model performance.

7.3 Future Work

Future study could build on this work by delving into a variety of key subjects. Using advanced augmentation techniques like Generative Adversarial Networks (GANs) and neural style transfer can result in more realistic augmented data. Cross-modal augmentation, which combines tactics from several data modalities, may increase model performance. Applying the developed methodology to real-world datasets from diverse domains would assist to validate its effectiveness and adaptability. Another interesting option is to

develop automated augmentation pipelines that select and apply the most relevant techniques based on data properties. Furthermore, training and evaluating models on these enriched datasets will be crucial for determining their practical utility and fine-tuning augmentation processes for maximum performance in a wide range of applications.

References

[1] "audio dataset". In: (). URL: https://www.zedge.net/ringtone/a895e626-2d4e-4706-996a-5b48852a4ac3.

[2] "Connor and Khoshgoftaar". In: (2022). URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0.

[3] "Devlin". In: (2019). URL: https://aclanthology.org/N19-1423/.

[4] "Junczys-Dowmunt". In: (2018). URL: https://aclanthology.org/N18-1055/.

[5] "Ravi". In: (2022). URL: https://arxiv.org/abs/2211.05047.

[6] "Sennrich". In: (2016). URL: https://aclanthology.org/P16-1009/.

[7] "sensor dataset". In: (). URL: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset.

[8] "text dataset". In: (). URL: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews?select=Reviews.csv.

[9] "Wang". In: (2022). URL: https://arxiv.org/pdf/2109.02054.

[10] "Wei and Zou". In: (2019). URL: %7Bhttps://aclanthology.org/D19-1670/%7D.