Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/51/20241354

Analysis of the methods and performances for data augmentation: Image, text and audio

Yuying Jin

Department of Computer Science, Penn State Behrend, Erie PA 16563, United State

yxj5181@psu.edu

Abstract. With the rapid development in computation ability as well as machine learning scenarios, various artificial intelligence applications can be achieved in recent years. With this in mind, this study will explore the application of data augmentation in machine learning and deep learning. To be specific, this paper first introduces the background and research history of data augmentation and then discusses the research progress in recent years. The basic description of this study describes the definition, common methods, and evaluation metrics of data augmentation in detail. At the same time, three data augmentation models, AutoAugment, AugGPT, and SpecAugment++, are introduced respectively, including their principles, experimental results, as well as evaluation. Finally, according to the analysis, the limitations and prospects of the field are discussed and demonstrated, as well as summarize the main findings and research implications of the full paper. Overall, these results shed light on guiding further exploration of data augmentation.

Keywords: Data augmentation, image data, text data, audio data.

1. Introduction

In the domain of deep learning, achieving decent performance for machine learning models hinges on the presence of a substantial quantity of training data; however, processes of data collecting and labelling are often pricy and time-consuming. To overcome these challenges, data augmentation (DA) stands out as a potent method, which can generate new data from existing data samples to raise the amount and variety of data while preserving the associated labels [1], making the model better generalize to unseen data. Across a spectrum of domains like image, text, and audio processing, DA methods have proven to be versatile and widely applicable [1-3]. Each of these data types, i.e., Image, text, and audio, have respective characteristics, requiring different data enhancement methods. For example, images can be enhanced using geometric transformations, colour space transformations, and noise injection [1]. Synonym replacement, deletion, random insertion, or swap are specific for text DA [2]. Audio can be enhanced using methods such as reverberation addition, noise addition, or pitch shift [3].

The earliest DA techniques can be traced back to simple transformations like horizontal flipping, colour space alterations, and random cropping [4]. The introduction of noise addition as a DA technique dates to the end of the 20th century [5]. In the early 2000s, SMOTE (Synthetic Minority Over-sampling Technique), a classic oversampling technology, was first proposed by Nitesh Chawla and others [6]. As computing power has advanced and deep learning has gained widespread popularity, more DA techniques have made significant progress in CV. In parallel, NLP and audio DA methods are growing,

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/51/20241354

yet challenges persist. Text DA poses challenges due to the discrete nature of text [2], and audio DA requires specialized knowledge [7].

The advanced DA methods includes following:

Image data

- Mixing images [1]: combines portions of different images to create novel images.
- Style Transfer [1]: allows to transfer of the visual style of an image to another, yielding stylized images with distinct artistic characteristics.
- Generative adversarial networks (GANs) [1]: generate realistic synthetic images keeping similar features with the original dataset.

Text data

- Data synthesis [8]: deletes, inserts, and replaces each token of corpus based on rules in random.
- Back translation [9]: translates text from one language to another and back introduces variations to create parallel training data.
- Word Embeddings [10]: uses pre-trained models, like Word2Vec, to substitute the original words in the text with synonyms or closely related terms

Audio data

- Sound synthesis: uses sound generation models like Variational Autoencoders (VAE) [11] to synthesize new sound samples.
- Random masking [12]: masks block of audio spectrogram in random.
- Mixing spectrogram [13]: masks spectrograms selected randomly and then mixes these spectrograms.

This article aims to explore recent prevalent data augmentation technologies in image, text, and audio domains, as well as present their main ideas, experiment results, and estimation, helping researchers enhance their application of these techniques to improve machine learning models.

2. Basic Descriptions

DA has been a technology widely used in deep learning. The main idea of DA is to produce additional or modified samples that are related to but slightly different from the original data via a set of transformations or perturbations to simulate more potential situations and variations in the real world. DA methods help models decline the risk of overfitting and adapt to unseen data. DA methods vary by application domain and data type. Here are some common DA methods:

Image data:

- Geometric transformation: includes operations such as translation, rotation, scaling, shearing, and flipping to change the angle, size, or position of the image.
- Colour Transform: adjust the colour, brightness, contrast, and saturation of the image to simulate different lighting conditions.
- Noise injection: adds random noise, such as salt and pepper noise, to the image.

Text data:

- Synonym replacement [10]: expands text data by replacing certain words or phrases in text with their synonyms.
- Deletion and insertion [10]: randomly delete words or phrases from the text and/or insert new words or phrases into the text to change the length and content of the text.
- Swapping order [10]: swaps the order of words or phrases in a text to introduce a different grammatical structure.

Audio data:

- Noise injection [14]: adds white noise or other types of noise to audio.
- Pitch Shift [15]: adjusts the pitch of audio but remains the duration same.
- Time Stretching [15]: modifies the speed of an audio sample with a specific ratio while keeping the same pitch.

DOI: 10.54254/2755-2721/51/20241354

When evaluating the effectiveness of data augmentation, since excessive use of augmented data can lead to issues with overfitting [16], researchers often use various metrics to measure improvements in model performance. Here are some common evaluation metrics:

- Accuracy: used for classification tasks, indicating the proportion of correctly classified samples.
- Recall: used for classification tasks, indicating the proportion of all real examples that are correctly classified.
- F1 Score: takes precision and recall into consideration.
- WER (Word Error Rate): used in speech recognition tasks to measure the difference between the recognition results and the reference text.

3. Models

3.1. AutoAugment–Automatic DA method

AutoAugment is widely used in both CV [17] and NLP [18] fields. This methodology represented a cutting-edge approach in the field of DA (2019), providing researchers and offers a powerful tool for improving model performance. AutoAugment uses a search algorithm like Reinforcement Learning to find optimal DA strategies automatically based on the dataset given. AutoAugment contains two components [17]. One is a search algorithm, and the other one is a search space. The search algorithm randomly selects DA policies, specifying types of operation, such as geometric and color transformations, along with the probability and magnitude of applying the operation. The search space has predefined sets of operations, probabilities, and magnitudes, representing potential DA strategies. Each set in this space represents a potential DA strategy. The search algorithm explores the search space, tries out different strategies, and evaluates the performance of each set of strategies with a cross-validation method. Then this algorithm will iteratively refine its hyperparameters and policies in the search space to identify the most effective strategy for the dataset. Essentially, the search space stores various sets of DA policies, while the search algorithm seeks out the most effective DA policies. Table 1 compares error rates (%) of different models with different DA methods on multiple datasets [17]. The DA methods tested include baseline (no DA methods), Cutout, and AutoAugment. AutoAugment can decrease the error rates of models. Each model can benefit from AutoAugment, especially PyramidNet+ShakeDrop and Shake-Shake (26 2x96d) models, whose error rates decreased significantly.

Table 1. Models and results.

Dataset	Model	Baseline	Cutout	AutoAugment
CIFAR-100	Wide-ResNet-28-10	18.8	18.4	17.1±0.3
	Shake-Shake(26 2x96d)	17.1	16.0	14.3±0.2
	PyramidNet-ShakeDrop	14.0	12.2	10.7±0.2
Reduced CIFAR-10	Wide-ResNet-28-10	18.8	16.5	14.1±0.3
	Shake-Shake(26 2x96d)	17.1	13.4	10.0±0.2
Reduced SVHN	Wide-ResNet-28-10	13.2	32.5	8.2±0.0
	Shake-Shake(26 2x96d)	12.3	24.2	5.9±0.0
CIFAR-10	Wide-ResNet-28-10	3.9	3.1	2.6±0.1
	Shake-Shake(26 2x96d)	2.9	2.6	2.0±0.1
	PyramidNet-ShakeDrop	2.7	2.3	1.5±0.1
SVHN	Wide-ResNet-28-10	1.5	1.3	1.1±0.0
	Shake-Shake(26 2x96d)	1.4	1.2	1.0±0.0

Table 2 showcases error rates (%) of a specific model trained with and without AutoAugment-transfer [17]. The development team investigates if AutoAugment is capable of transferring augmentation policies across datasets, and this method is called AutoAugment-transfer. One can see that AutoAugment-transfer decreases error rates significantly on various datasets even with small sizes.

DOI: 10.54254/2755-2721/51/20241354

AutoAugment has significantly achieved state-of-the-art (SOTA) precision across various datasets, including CIFAR-100, Reduced CIFAR-10, Reduced SVHN, CIFAR-10, and SVHN. The AutoAugment development team also investigated how DA policies searched may be shared among datasets, referred to as "transferability", aiming to reduce the need for its resource-intensive policy searches. The result shows AutoAugment has remarkable transferability of DA policies. ImageNet-derived DA policies consistently improve generalization accuracy on FGVC datasets. The substantial performance improvements across different datasets showcased the ability of AutoAugment to learn and apply data augmentation policies, resulting in performance boosts across diverse datasets, even when dealing with smaller datasets. AutoAugment undoubtedly is a milestone of DA method development. Still, its search algorithm is resource-intensive. Some researchers have developed other DA methods based on AutoAugment to improve time efficiency, such as Adversarial AutoAugment [19] and Fast AutoAugment [20].

Table 2. Error rates (%) of a specific model trained with and without AutoAugment-transfer

Dataset	Train Size	Classes	Baseline	AutoAugment-transfer
Caltech-101	3.060	102	19.4	13.1
Oxford-IIIT Pets	3,680	37	13.5	11.0
FGVC Aircraft	6,667	100	9.1	7.3
Stanford Cars	8,144	196	6.4	5.2
Oxford 102 Flowers	2,040	102	6.7	4.6

Table 3. Accuracy of different DA methods

DA Methods	Amazon		Symptoms		PubMed20K	
	BERT	BERT C	BERT	BERT C	BERT	BERT C
Raw	.734	.745	.636	.606	.792	.798
BackTranslationAug	.757	.748	.778	.747	.812	.83
ContextualWordAugUsingBert(Insert)	.761	.750	.697	.677	.802	.811
ContextualWordAugUsingBert(Substitute)	.770	.757	.626	.667	.815	.830
ContextualWordAugUsingDistilBERT(Insert)	.759	.762	.707	.747	.796	.796
ContextualWordAugUsingDistilBERT(Substitut	.787	.766	.667	.646	.797	.800
ContextualWordAugUsingRoBERTA(Insert)	.775	.768	.758	.707	.815	.814
ContextualWordAugUsingRoBERTA(Substitute	.745	.730	.727	.667	.782	.782
CounterFittedEmbeddingAug	.754	.741	.667	.626	.805	.805
InsertCharAugmentation	.771	.775	.404	.475	.826	.831
InsertWordByGoogleNewsEmbeddings	.816	.794	.636	.677	.786	.784
KeyboardAugmentation	.764	.766	.545	.505	.809	.815
OCRAugmentation	.775	.782	.768	.778	.789	.789
PPDBSynonymAug	.691	.690	.697	.758	.795	.829
SpellingAugmentation	.727	.736	.697	.707	.808	.811
SubstituteCharAugmentation	.762	.768	.535	.586	.816	.821
SubstituteWordByGoogleNewsEmbeddings	.729	.741	.727	.727	.807	.822
SwapCharAugmentation	.762	.766	.475	.485	.797	.801
SwapWordAug	.771	.766	.687	.727	.798	.794
WordNetSynonymAug	.805	.798	.616	.758	.761	.757
ChatGPT (2-shot)	.753		.98		.748	
AugGPT	.816	.826	.889	.899	.835	.835

DOI: 10.54254/2755-2721/51/20241354

3.2. AugGPT-Text DA method

AugGPT is a text DA method based on ChatGPT. Its main idea involves transforming input text into several variations that convey similar underlying conceptions but in slightly different ways [21]. AugGPT continuously performs a SOTA performance in text DA method as of 2023. During pre-training, AugGPT employs transformer blocks to extract data features and learn how to effectively predict the subsequent token within a sequence. Then, AugGPT applies Reinforcement the Learning from Human Feedback (RLHF) technique to fine-tune its pre-training language model [21]. RLHF consists of three steps: Supervised Fine-tuning (SFT), Reward Modeling (RM), and Reinforcement Learning (RL) [21]. In the phase of SFT, ChatGPT learns from the questions and responses provided by humans to make its answering strategy more like human responses. RM assigns a score to each pair of questions and answers, letting the ChatGPT get better at judging responses. Through RL, ChatGPT continues its training with the feedback it receives, furthering its language capabilities. In the conducted experiments, the development team used datasets Amazon, Symptoms, and PubMed20K to compare various DA methods. Table 3 displays the accuracy of different DA methods. AugGPT achieves optimal accuracies within these three datasets. In the dataset of Amazon, both AugGPT and InsertWordByGoogleNewsEmbeddings are the top performers. In the Symptoms dataset, AugGPT boosts accuracy from 63.6% to 89.9%. Similarly, in PubMed20K, AugGPT presents a significant rise, reaching 83.5% compared to the baseline accuracy of 79.2%. These results represent the high effectiveness of AugGPT in enhancing the performance of different machine-learning models across

AugGPT performs well on different datasets and has significant performance improvement capabilities. Its remarkable results on various machine learning models and diverse datasets demonstrate the high effectiveness of AugGPT in enhancing model performance in different applications. This result can be attributed to AugGPT's extensive pre-training and effective data augmentation strategies. The limitation of AugGPT is it might generate invalid or incorrect augmented data if ChatGPT does not possess such domain-specific knowledge or does not understand the context. Besides, AugGPT may generate data that reflects stereotypes or biases present in the training data like ChatGPT [22].

3.3. SpecAugment++–Audio DA method

SpecAugment++ [23] is inspired by SpecAugment [24]. Both SpecAugment++ and SpecAugment are DA methods for speech recognition. Instead of using the original audio, SpecAugment modifies the log mel spectrogram by applying DA methods like time wrapping, frequency masking, and time masking. SpecAugment++ focuses on the two masking techniques, frequency and time masking, applying them to the hidden states. Additionally, SpecAugment++ explores three approaches for masking: zero-value masking (ZM), mixture masking (MM), and cutting masking (CM). ZM substitutes consecutive time and frequency channels with zeros directly. MM and CM involve combining time frames and frequency channels from other samples contained in the same mini-batch.

Table 4. Comparison of classification accuracy (%) among various DA methods using datasets DCASE 18 and 19.

DA Methods	DCASE 18	DCASE 19
No augmentation	74.3±.59	78.9±.80
Mixup (2017)	75.5±.62	79.3±.71
SpecAugment (2019)	74.9±.81	79.1±1.05
BC Learning (2017)	75.8±.66	8.0±.76
SpeechMix (2020)	75.8±.48	8.7±.69
SpecAugment++ (ZM)	76.2±.59	8.6±.82
SpecAugment++ (CM)	76.9±.73	81.4±.94
SpecAugment++ (MM)	77.0±.52	82.6±.66

DOI: 10.54254/2755-2721/51/20241354

Table 4 shows the comparison of classification accuracy (%) among various DA methods using datasets DCASE 18 and 19. This comparison also includes the result of three different approaches of SpecAugment++. SpecAugment++ overperforms other prevalent DA methods such as mixup, SpecAugment, BC learning, and Speech Mix. Among the three different approaches of SpecAugment++, ZM is slightly inferior compared to MM and CM, with MM performing best on both datasets with accuracy of around 77% and 82.6% separately. Fig. 1 presents how accuracy (%) changes with different masking ratios (%) for both time and frequency, using three various masking approaches, CM, MM, and ZM. Among these approaches, MM performs better than the other two [23]. The accuracy of all three approaches improves initially and then decreases from 25% as the ratio for time or frequency increases. When the ratio is larger than 40%, the performance starts to decline significantly. SpecAugment++ has shown significant promise in improving ASR model performance, consistently outperforming other popular DA methods. Its MM approach performs the best on various datasets while ZM does not demonstrate good enough generalization capability, which highlights the importance of choosing a suitable masking strategy in achieving optimal results with SpecAugment++. In addition, these masking approaches are sensitive to masking ratios, which means choosing the right hyperparameter is critical for optimal performance. This sensitivity to masking ratios underscores the need for careful parameter tuning when applying SpecAugment++.

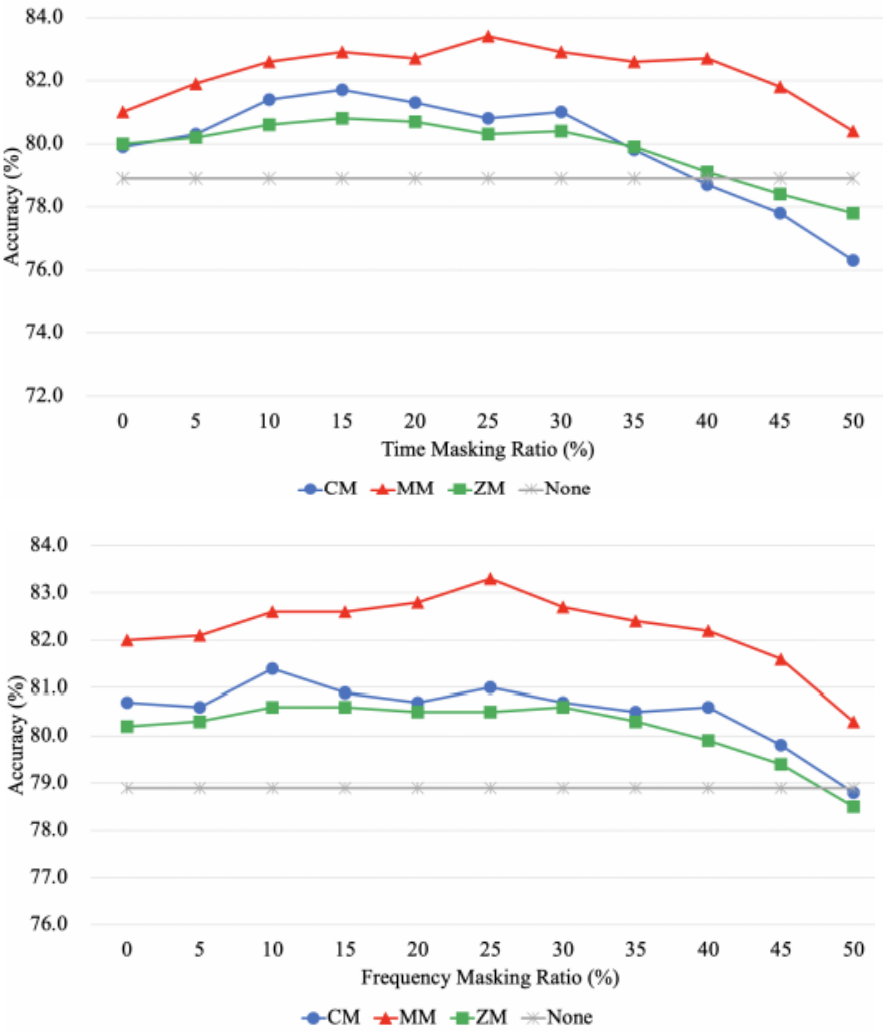

Figure 1. Accuracy as a function of time (upper) and frequency masking ratio (lower) [23].

DOI: 10.54254/2755-2721/51/20241354

4. Limitations and prospects

Many DA methods are domain-specific, making it difficult to generalize their application to different machine-learning tasks. Each machine learning domain has its unique data characteristics. For example, shearing and rotating the spectrum of audio data could distort the data while they play an important role in the CV domain. In contrast, NLP tasks involve sequences of words, making operations like rotation irrelevant. Secondly, DA methods may lead to overfitting. If the augmentation policies applied are invalid, they might introduce noise or inconsistency, changing the data feature distribution too much. Researchers are required to carefully consider appropriate DA strategies. Some DA models are hyperparameter sensitive. This sensitivity may lead to large fluctuations in model performance on various datasets, requiring fine-tuning, which increases the complexity of using these DA methods. Furthermore, computing cost is another significant limitation, especially in large-scale machine learning applications, using DA methods may result in expensive computing resource requirements. Despite the limitations of DA methods, the prospects are promising. As computing power increases continuously, the cost issue will gradually diminish. Future research may focus on developing more robust DA methods that are less sensitive to hyperparameters. Adaptive augmentation strategies like AutoAugment might be developed and could find broader applications across various domains including image, text, and audio data processing.

5. Conclusion

To sum up, this study has presented DA methods in image, text, and audio. Three prominent DA methods were discussed: AutoAugment, an automatic DA method in image and text, can search optimal DA policies and presents remarkable policy transferability across datasets. AugGPT is a SOTA text DA method based on ChatGPT. SpecAugment++ focuses on audio DA and presents superior performance compared to other popular audio DA methods, especially with its MM approach. However, it is essential to fine-tune masking ratios for optimal results. The limitations of DA methods include domain-specific applicability, potential overfitting, sensitivity to hyperparameters, and increased computational costs. As the computing power increases and more advanced DA methods develop, the potential for improving machine learning models through DA techniques will become increasingly evident.

References

[1] Khosla C and Saini B S 2020 International Conference on Intelligent Engineering and Management (ICIEM) pp 79-85.
[2] Feng S Y, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T and Hovy E 2021 arXiv preprint arXiv:2105.03075.
[3] Nam H, Kim S H and Park Y H 2022 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) pp 4308-4312.
[4] Shorten C and Khoshgoftaar T M 2019 J Big Data vol 6(1) p 6.
[5] Holmstrom L and Koistinen P 1992 IEEE Transactions on Neural Networks vol 3(1) pp 24-38.
[6] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 arXiv preprint arXiv:11061813.
[7] Wei S, Zou S, Liao F and Lang W 2020 Journal of Physics: Conference Series vol 1453(1) p 012085.
[8] Yang L, Wang C, Chen Y, Du Y and Yang E 2019 arXiv preprint arXiv:190913302.
[9] Sennrich R, Haddow B and Birch A 2016 Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics vol 1 pp 86–96.
[10] Li B, Hou Y and Che W 2022 AI Open vol 3 pp 71-9.
[11] Saldanha J, Chakraborty S, Patil S, Kotecha K, Kumar S and Nayyar A 2022 PLoS ONE vol 17(8) p e0266467.
[12] Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D and Le Q V (201 arXiv preprint arXiv:190408779.
[13] Kim G, Han D K and Ko H 2021 arXiv preprint arXiv:21080302.

DOI: 10.54254/2755-2721/51/20241354

[14] Abayomi-Alli O O, Damaševičius R, Qazi A, Adedoyin-Olowe M and Misra S 2022 Electronics vol 11(22) p 3795.
[15] Ferreira-Paiva L, Alfaro-Espinoza E, Almeida V M, Felix L B and Neves R V 2022 XXIV Brazilian Congress of Automatics (CBA) pp 122-127.
[16] Chen T and Nguyen-Thi T A 2021 Computational Social Networks vol 8(1) p 1.
[17] Cubuk E D, Zoph B, Mané D, Vasudevan V and Le Q V 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp 113-123/
[18] Niu T and Bansal M 2019 arXiv preprint arXiv:190912868.
[19] Zhang X, Wang Q, Zhang J and Zhong Z 2019 arXiv preprint arXiv:191211188.
[20] Lim S, Kim I, Kim T, Kim C and Kim S 2019 Neural Information Processing Systems vol 32.
[21] Dai H, Liu Z, Liao W, Huang X, Cao Y, Wu Z and Li X 2023 arXiv preprint arXiv:230213007.
[22] Laskar M T R, Bari M S, Rahman M, Bhuiyan M A H, Joty S and Huang J X 2023 arXiv preprint arXiv:230518486.
[23] Wang H, Zou Y and Wang W 2021 arXiv preprint arXiv:210316858.
[24] Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D and Le Q V 2019 arXiv preprint arXiv:190408779.