ENHANCING MACHINE LEARNING MODEL PERFORMANCE THROUGH DATA AUGMENTATION TECHNIQUES ACROSS VARIED DATASET SIZES

Maxim PLĂMĂDEALĂ[1], Eduard BALAMATIUC[2*], Marin NEGAI[3], Cristofor FIȘTIC[4]

[1]*Department of Software Engineering and Automatics, gr. FAF-222, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*
[2]*Department of Software Engineering and Automatics, gr. FAF-221, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*
[3]*Department of Software Engineering and Automatics, gr. FAF-223, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*
[4]*Department of Software Engineering and Automatics, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*

*Corresponding author: Balamatiuc Eduard, eduard.balamatiuc@isa.utm.md

Scientific coordinator: **Dumitru CIORBĂ**, conf. univ., dr.

Abstract. *In the realm of machine learning, the challenge of limited data availability often hampers the development and performance of predictive models. Data augmentation, the process of artificially expanding a dataset through various modifications and transformations, presents a promising avenue to mitigate these limitations. This article embarks on a theoretical exploration of data augmentation techniques and their potential to bolster the effectiveness of machine learning models, irrespective of the initial dataset size. The core argument posits that data augmentation can serve as a critical tool in enhancing model performance, particularly when confronted with sparse data. It emphasizes the need for a thoughtful selection of augmentation techniques that align with the characteristics of the data and the objectives of the machine learning task at hand. Furthermore, the abstract posits a theoretical framework for understanding the relationship between dataset size and the efficacy of data augmentation, suggesting that the impact of augmentation might vary across different data scales and model complexities. In sum, this article aims to shed light on the strategic importance of data augmentation in the field of machine learning, advocating for its consideration as an essential component in the model development process, especially in scenarios characterized by data scarcity.*

Keywords: *data science, data augmentation, machine learning.*

Introduction

Machine learning is a cornerstone of many technological advancements, driving innovations in areas like computer vision, natural language processing, and recommendation systems. However, the performance of these models heavily relies on the quality and quantity of data used for training. Limited data availability can lead to suboptimal model performance, hindering their ability to generalize effectively to unseen data [1].

Data augmentation emerges as a powerful technique to address this challenge. By artificially creating new variations of existing data points, data augmentation expands the training set, fostering model robustness and generalizability. This article delves into the theoretical underpinnings of data augmentation and its potential to enhance model performance across varying dataset sizes. Data augmentation serves as a critical tool, particularly in scenarios with limited data, by effectively mitigating the effects of data scarcity. However, the success of data augmentation hinges on the judicious selection of techniques tailored to the specific data characteristics and the learning task at hand. By investigating how augmentation impacts models across different data scales and complexities, this article aims to illuminate its strategic importance

in the machine learning landscape. Ultimately, data augmentation becomes an essential component of the model development process, especially when dealing with limited data resources.

Background and Related Work

Machine learning algorithms learn patterns from data to make predictions on unseen examples. Supervised learning, a prevalent paradigm, utilizes labeled data for training. The model's performance hinges on its ability to generalize effectively to new data, which can be hampered by limited training data. This phenomenon, known as data scarcity, often leads to overfitting, where the model memorizes the training data peculiarities and fails to adapt to unseen scenarios. To reduce this phenomenon, multiple techniques were proposed. One of them was regularization that adds constraints and penalties to the model parameters to prevent it become too complex. Another technique was batch normalization [2] which is used in neural networks that involves the activations of each layer by adjusting and scaling them to have a mean of zero and a standard deviation of one [3]. Another salvation for the low accuracy and overfitting, could be transfer learning and fine-tuning that are allowing to retrain the model for a new problem, or add up tune the data to more data.

Data augmentation in this case tackles the challenge of data scarcity by artificially expanding the training dataset. This technique involves applying various transformations and modifications to existing data points, generating new variations that retain the core information [4]. In fact, data augmentation is not new, however its applicability is very extensive and can be helpful not only for computer models but for real-world tasks [3].

Data augmentation can be variate. It takes form of geometric transformations like flips, rotations, scaling or color space manipulations like modifying brightness or contrast. Mathematically speaking, these are affine transformations and are related to images and image-like data. Generally, this type of augmentation can be resumed to this formula (Eq. (1)):

$$y = Wx + b \tag{1}$$

in the formula (1), y represents the transformed image, W is the matrix representing the linear transformation applied to the original image data, x is the vector of the original image data, and b is the vector of the bias terms that translates the image. This type of transformations are proving themselves best, as they generate new data, thus making the model not too complex and trained for new test data.

Other tools that can provide data augmentation are Generative Adversarial Networks (GANs). GANs have shown remarkable capabilities in generating realistic data samples, especially in image synthesis tasks. For instance, CycleGAN, a variant of GAN, can perform style transfer between images from different domains without paired training data, opening up possibilities for various applications in art, fashion, and design.

Integrating GANs with data augmentation techniques presents a promising approach to address data scarcity issues. By leveraging GANs to generate synthetic data, it’s possible to effectively expand the training dataset, enhancing the model's robustness and generalization capabilities.

In the next section, more attention will be focused on techniques, with a detailed exploration of both their mathematical principles and practical applications.

Data Augmentation Techniques

Numerical Data Augmentation

Numerical data augmentation involves manipulating the numerical features in a dataset to generate synthetic data points that retain the statistical properties of the original data:

- Noise injection happens when small, random variations are added to numerical values, typically drawn from a distribution like Gaussian or uniform, to simulate variability.

This is mathematically represented as $X_{aug} = X + \varepsilon$, where X is the original data and ε is the random noise.

- Scaling and shifting involve linear transformations of the form $X_{aug} = aX + b$, where a and b are scaling and shifting parameters, respectively. This can help in simulating different operational conditions in scenarios like sensor data monitoring [5].

Categorical Data Augmentation

Categorical data augmentation is more complex due to the discrete nature of the data:

- Smoothing techniques, like additive or Laplace smoothing, to manage the distribution of categorical variables, particularly or underrepresented categories.
- Category embedding and swapping, which involves representing categories in a continuous vector space and ten performing operations similar to numerical data augmentation within this space [6].

Textual Data Augmentation

Textual data augmentation helps to increase the robustness of natural language processing (NLP) models by introducing variety in the training data, which aids in learning more general patterns and reducing overfitting.

- Synonym Replacement: this technique involves replacing words in a sentence with their synonyms to create a new sentence with the same meaning but different wording. It helps the model learn that different words can express similar meanings, enhancing its ability to understand context and semantics. For example, the sentence "The quick brown fox jumps over the lazy dog" can be altered to "The fast brown fox leaps over the sluggish dog."
- Back-translation: this involves translating a sentence into a foreign language and then translating it back to the original language. The process often introduces linguistic nuances and variations that wouldn't be present in the original text, thus enriching the language model's training data. For instance, the English sentence "He goes to school by bus" might be translated to French and back to English as "He takes the bus to go to school," introducing a different sentence structure [7].
- Sentence Shuffling: by rearranging the order of sentences or phrases within a text, this method enhances the model's ability to understand and predict context and coherence in language. However, it's important to maintain a logical flow of information, so this technique is more suitable for texts where the order of information is not critical to understanding.

Temporal Data Augmentation

Temporal data augmentation is used in time-series analysis to improve the prediction accuracy of models by training them on varied time-based scenarios.

- Time Warping: similar to stretching or compressing a timeline, time warping applies nonlinear transformations to the temporal axis of the data. This simulates varying speeds of event occurrences, helping models to better learn and predict temporal dynamics under different conditions [7].
- Window Slicing: by cutting the time series into different windows or segments, this technique allows models to train on diverse temporal slices, thus improving their generalization across time. It's especially useful for detecting patterns or anomalies in time-series data that occur over inconsistent time intervals.

Generative Data Augmentation

Generative models like GANs and VAEs create new data samples that can augment existing datasets, especially in areas where data collection is challenging.

- GANs for Realistic Sample Generation: GANs can generate highly realistic samples by learning the distribution of the original data. They are particularly useful in generating complex data like images or sounds, where they can create new instances that are hard to distinguish from real ones, thus providing additional training material for models [6].
- VAEs for Data Interpolation: Variational Autoencoders (VAEs) generate new data points by interpolating in the latent space, a compressed representation of the data. This process allows for the generation of new samples that are variations of the original data, maintaining the core statistical and structural properties. It's beneficial for enhancing the diversity of datasets without straying from their inherent characteristics [7].

Practical Applications in Tabular Data

In practical scenarios, such as credit risk assessment or customer churn prediction, data augmentation can be pivotal. For example:

- Credit scoring: augmenting financial datasets by injecting noise into numerical features like income or loan amount can help in creating more robust models that are less sensitive to small variations, reflecting real-world uncertainties.
- Customer churn analysis: for categorical features like subscription type or service plan, techniques like category swapping (where similar categories are interchanged) can simulate various customer behavior patterns, leading to a model that better generalizes across customer segments.

Challenges and Considerations

When augmenting tabular data, several challenges need to be addressed:

- Preserving data integrity: it's crucial to ensure that augmented data still respects the intrinsic relationships and constraints within the original dataset, such as the logical correlation between features.
- Feature importance: the augmentation process should consider the relative importance of features, focusing more on those that have a significant impact on the model's predictions.
- Over-augmentation: excessive augmentation can lead to model training on essentially synthetic data that may not represent real-world scenarios, potentially harming the model's ability to generalize [8].

Data Augmentation across different dataset sizes

As said, data augmentation serves as a pivotal strategy in enhancing model generalization, particularly when dealing with limited data volumes. Various studies underscore its efficacy across dataset sizes, highlighting nuanced impacts.

Small Datasets: in scenarios with scant data, augmentation emerges as a pivotal tool to curb overfitting [9]. Research by Shorten and Khoshgoftaar [10] emphasizes its role in expanding datasets, mitigating overfitting risks, and improving model performance. Techniques like random rotations, flipping, and cropping diversify training samples, facilitating robust feature extraction [11].

Medium Datasets: while augmentation remains beneficial in medium-sized datasets, its impact may moderate due to the sufficient data volume available [9]. However, tailored augmentation techniques remain pivotal. For instance, research by Cubuk et al. [10] underscores

the significance of domain-specific augmentation strategies for improved performance in image classification tasks.

Large Datasets: in large datasets, the utility of data augmentation varies based on task complexity [10]. While ample data may reduce augmentation's impact, it remains vital for tasks requiring nuanced pattern recognition [11]. Notably, in medical imaging, augmentation aids in capturing subtle variations crucial for accurate diagnoses [9].

Very Large Datasets (Big Data): extremely large datasets pose unique challenges, where the marginal utility of augmentation may diminish [10]. Nevertheless, meticulously crafted augmentation strategies could still benefit specialized tasks [11].

Understanding these nuances aids in crafting effective augmentation strategies for optimal model performance.

Conclusions

In summary, data augmentation serves as a fundamental technique in machine learning, providing crucial solutions to the persistent challenge of limited data availability. This article has delved into the diverse landscape of augmentation techniques, spanning numerical, categorical, textual, temporal, and generative domains, elucidating their pivotal role in enhancing model robustness and generalization.

By exploring the theoretical underpinnings and practical applications of augmentation across different dataset sizes, this discussion has underscored its strategic significance in curbing overfitting and fostering model adaptability. From injecting noise in numerical data to performing category embedding in categorical data and employing sophisticated methods like back-translation in textual data, augmentation techniques offer versatile solutions tailored to diverse data modalities.

However, the successful implementation of data augmentation requires navigating various challenges, including preserving data integrity, considering feature importance, and mitigating the risk of over-augmentation. Despite these challenges, augmentation proves invaluable across different dataset scales, with nuanced impacts observed in small, medium, large, and very large datasets.

As the machine learning landscape continues to evolve, with advancements in generative models like GANs and VAEs expanding the horizons of data augmentation, understanding these nuances becomes increasingly crucial. By equipping practitioners with insights into the intricacies of augmentation techniques and their impacts across varying data scales, this article aims to empower the development of effective augmentation strategies, thereby maximizing model performance and driving advancements in predictive modeling across a myriad of domains.

References

[1] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, p. 60, 2019. doi: 10.1186/s40537-019-0197-0.

[2] H. Zhang et al., “Mixup: Beyond Empirical Risk Minimization,” in International Conference on Learning Representations, 2018.

[3] Y. Jiang et al., “Smart Augmentation Learning an Optimal Data Augmentation Strategy,” arXiv:1703.03069, 2017.

[4] B. Wang and D. Klabjan, “Regularization for unsupervised deep neural nets,” CoRR, abs/1608.04426, 2016.

[5] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. International Conference on Machine Learning, PMLR, 2015, pp. 448-456.

[6] J. Wang and L. Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," arXiv:1712.04621v1 [cs.CV], Dec. 13, 2017.
[7] E. D. Cubuk et al., "AutoAugment: Learning Augmentation Policies from Data," arXiv:1805.09501, 2018.
[8] J. Choi et al., "Deep Learning in Medical Imaging: General Overview," Korean J. Radiol., vol. 19, no. 4, pp. 570-584, 2018.
[9] J. Pérez and B. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," arXiv:1712.04621, Dec. 2017.
[10] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, pp. 60, 2019.
[11] A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, 2012, pp. 1097-1105.