Data augmentation: A literature review

Joao Fonseca[1]*, Fernando Bacao[1]

[1]NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author

Postal Address: NOVA Information Management School, Campus de Campolide, 1070–312 Lisboa, Portugal

Telephone: +351 21 382 8610

This is the abstract for a literature review

1 Introduction

The performance of Machine Learning models is highly dependent on the quality of the training dataset used [1, 2]. The presence of imbalanced and/or small datasets, target labels incorrectly assigned, outliers and high dimensional input spaces reduce the prospects of a successful machine learning (ML) model implementation [2, 3, 4]. In particular, deep learning architectures are often limited by a natural inclination to overfitting, label noise memorization and catastrophic forgetting [5].

2 Brief Historical Perspective

3 Data Augmentation Taxonomy

4 Review of the State-of-the-art

5 Algorithmic applications

References

[1] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma, “Data set quality in machine learning: Consistency measure based on group decision making,” *Applied Soft Computing*, vol. 106, p. 107366, 2021.

[2] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[3] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[4] S. Salman and X. Liu, “Overfitting mechanism and avoidance in deep neural networks,” *arXiv preprint arXiv:1901.06566*, 2019.

[5] Z. Xie, F. He, S. Fu, I. Sato, D. Tao, and M. Sugiyama, “Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting,” *Neural computation*, vol. 33, no. 8, pp. 2163–2192, 2021.