Research Trends and Applications of Data Augmentation Algorithms

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

Over past years, researchers have developed complex deep learning classifiers, which typically require a tremendous amount of data and computational power to train. Although, these resources are not always accessible to most organizations and/or practitioners. Concurrently, current and past research have shown that even simple classification algorithms can reach state-of-the-art performance on computer vision tasks given a robust method to artificially augment the training dataset. Because of this, data augmentation became a popular topic in the past few years, particularly among computer vision researchers. In this paper we identify the main areas of application of data augmentation algorithms, the main types of algorithms used, significant research trends and their progression over time and research gaps in data augmentation literature. This was done through using a network-based, text mining-based and exploratory analysis of literature collected through the Scopus database. We expect readers to understand the potential of data augmentation as well as identify future research directions and open questions within data augmentation research.

1 Introduction

Machine Learning models are highly dependent on the quality of the training dataset used [1]. The presence of imbalanced and/or small datasets, target labels incorrectly assigned, outliers and high dimensional input spaces may significantly reduce the prospects of a successful machine learning model implementation. Although the quality of all classifiers are affected by a small training dataset, Deep Learning classifiers are particularly sensitive to it. Depending on the complexity of the model, deep learning models show a high degree of performance variability even with large amounts of training data, particularly when the model is tested over a fully independent validation/test set [2]. Various problems contribute towards this performance variability, *e.g.,* biased training dataset, varying conditions of data collection across test/train set and small training data.

When the training set contains significant limitations, the trained model model runs into the problem of overfitting. There are different strategies to reduce overfitting, known as regularization methods, such as pruning, early stopping, dropout, batch normalization, transfer learning, pretraining, one-shot and

zero-shot learning, and data augmentation [3]. The later, data augmentation, are techniques used to increase the diversity of data in a training dataset through the production of artificial observations.

Depending on the generative model used, data augmentation may be used to various types of problems and data. In 2011, Jürgen Schmidhuber's group showed that a MLP ensemble architecture can achieve state-of-the-art performance on computer vision benchmarks given strong enough data augmentation [4, 5]. Although state-of-the-art research has progressed significantly since then, two recent papers developed by Google Brain and Facebook research teams show that Schmidhuber's group's findings still hold true. Specifically, in [6, 7] the authors discuss two similar MLP ensemble architectures, showing that the proposed model attains a comparable performance to convolutional neural networks and attention-based networks. Another recent study also discusses a related MLP architecture with similar findings, while suggesting that the strong performance of computer vision models may be attributable mainly to the inducive bias produced by the patch embedding and the carefully-curated set of training augmentations [8].

Research on data augmentation methods has gained significant popularity in recent years. As such, there were some efforts in the past to establish a taxonomy and distinction of the different types of data augmentations methods [3], although outdated/incomplete. The most cited literature review on data augmentation was focused on image data augmentation for deep learning in 2019 [3]. Since then, research on data augmentation methods have progressed significantly. In this paper we focus on current and past research trends of data augmentation methods, its different applications and use cases. This was done with an extensive analysis of the title, keywords and abstract of a large set of literature related to data augmentation, collected through the Scopus database. The analysis was done in three phases. We started by performing an exploratory data analysis to identify the most significant literature, journals and conferences within the field of data augmentation. Then, we analyse the articles' author keywords by constructing a network and extracting and identifying communities of keywords. Finally we used a text mining approach to extract additional applications and methods using the articles' abstracts, as well as validate the findings discussed with the keyword analysis.

This paper is structured as follows: Section 2 describes the procedures defined throughout the different analyses. Section 3 presents and discusses the findings drawn from the analyses, as well as research gaps and open questions in data augmentation research. Section 4 summarizes the main findings discussed throughout the study.

2 Methodology

In this section we describe the procedures defined for the literature collection, data preprocessing and literature analysis. The analysis of the literature was developed with 3 different approaches. Throughout the analyses, data preprocessing and hyperparameter tuning was developed iteratively. The procedure adopted in this manuscript is shown in Figure 1.

The literature collection procedure is described in Subsection 2.1. The data and text preprocessing is described in Subsection 2.2. The exploratory data analysis described in Subsection 2.3 was done to understand which manuscripts, journals and conferences are most significant within the field of Data Augmentation. The manuscripts' keywords were used to construct a network of keywords (described in Subsection 2.4) and study the different communities of keywords found in the network. The topic

modelling and parameter tunning is described in Subsection 2.5. The abstract embeddings procedure is described in Subsection 2.6.

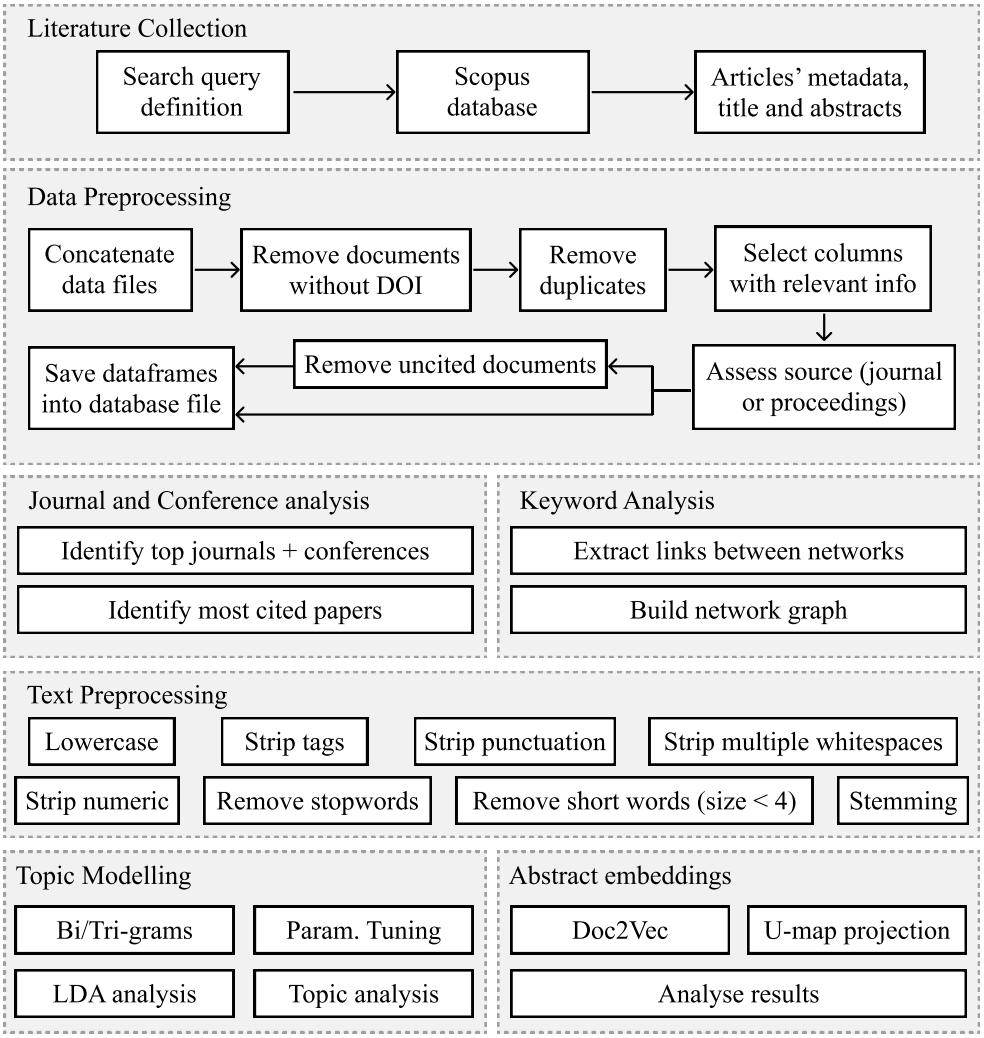

Figure 1: Diagram of the proposed literature analysis approach.

2.1 Literature Collection

The focus of this literature analysis is to understand the different algorithms, domains and/or tasks that employ data augmentation techniques. Therefore, we search for documents containing the keyword "data augmentation" in the search query. The results were then limited to conference papers and journal articles written in English that were published in the past 15 years. Due to the large amount of results found, using solely the Scopus database was found to be sufficient. One of the goals during the search query design was to come up with a simple and unbiased query. The resulting query is shown below:

```
KEY ( "data augmentation" )  AND  ( LIMIT-TO ( LANGUAGE ,  "English" ) )
AND ( LIMIT-TO ( DOCTYPE ,  "cp" )  OR  LIMIT-TO ( DOCTYPE ,  "ar" ) )
```

```
AND (
        LIMIT-TO ( PUBYEAR ,  2021 )  OR  LIMIT-TO ( PUBYEAR ,  2020 )
    OR  LIMIT-TO ( PUBYEAR ,  2019 )  OR  LIMIT-TO ( PUBYEAR ,  2018 )
    OR  LIMIT-TO ( PUBYEAR ,  2017 )  OR  LIMIT-TO ( PUBYEAR ,  2016 )
    OR  LIMIT-TO ( PUBYEAR ,  2015 )  OR  LIMIT-TO ( PUBYEAR ,  2014 )
    OR  LIMIT-TO ( PUBYEAR ,  2013 )  OR  LIMIT-TO ( PUBYEAR ,  2012 )
    OR  LIMIT-TO ( PUBYEAR ,  2011 )  OR  LIMIT-TO ( PUBYEAR ,  2010 )
    OR  LIMIT-TO ( PUBYEAR ,  2009 )  OR  LIMIT-TO ( PUBYEAR ,  2008 )
    OR  LIMIT-TO ( PUBYEAR ,  2007 )  OR  LIMIT-TO ( PUBYEAR ,  2006 )
)
```

The search query resulted in 4281 documents. The resulting data selection/filtering pipeline is shown in Figure 2. Due to the limitations in the Scopus data export (maximum 2000 documents per export), the data was split in four different time periods and exported separately: 2006 until 2018, 2019, 2020 and 2021, which produced four CSV files.

2.2 Data Preprocessing

The data preprocessing stage and amount of documents dropped is represented in Figure 2. The data was first concatenated into a single data frame. During this process, we found that one of the exported references had a corrupted line, which caused the loss of one additional document. Since the DOI can be used as a unique identifier for intellectual property [9], references without a DOI were disregarded from further analysis, while the ones with the same identifiers are removed (*i.e.*, only one of the repeating entries is kept).

This dataset was kept to perform the analysis described in Subsection 2.3. However, further preprocessing was done for the remaining parts of the literature analysis. References without any citations were excluded for the keyword network and topic modelling analyses. Finally, only the documents containing keywords in Scopus' database were used to prepare the network analysis.

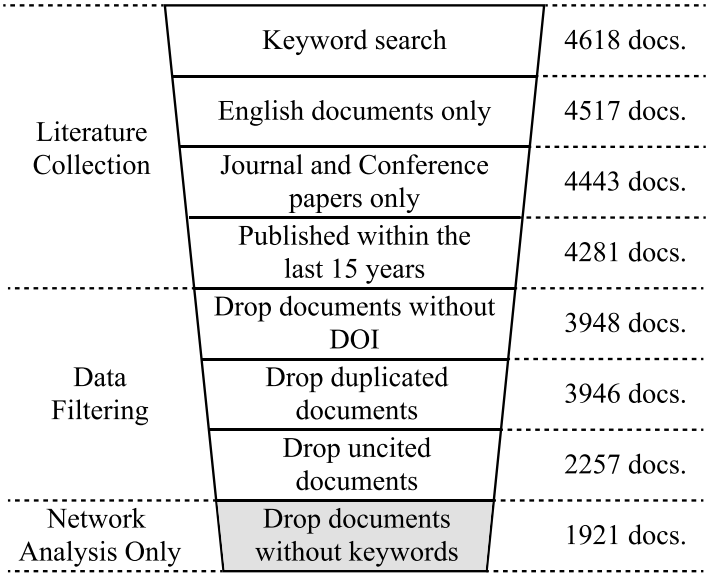

Figure 2: Data filtering pipeline.

2.3 Journal and Conference analysis

The exploratory analysis developed on the preprocessed dataset was targeted towards the identification of the most significant works, journals and conferences. We used the citation count as a proxy to understand the impact of a specific manuscript within the research community.

The identification of the most significant conferences and journals is done by sorting each type of publication according to the number of citations per document. Conferences and journals with less than 10 papers published in the area are not considered in this analysis.

2.4 Keyword Analysis

The analysis of keywords is expected to uncover general trends in data augmentation research and its applications. The keyword "data augmentation" was removed since it would link with all other keywords. Keywords are connected based on their co-occurrence in each research paper to form the edges of the network. It consists of an undirected graph whose weights are based on the total citation count for the papers containing a given keyword pair and is calculated as weight $= \log(\text{citations}) + 1$ to avoid a potential bias caused by highly cited research articles. The size of the nodes were determined with a logarithmic transformation of each node's page rank.

Keyword combinations showing up in only one document are removed from further analysis. The keyword network is then analysed using Python and the communities were found using the greedy modularity

maximization algorithm proposed in [10]. The results of the analysis and community detection were ported to Gephi to produce the final visualizations.

2.5 Topic Modelling

The extraction of topics was done using the publication's abstracts. The words were tokenized and all tags, special characters, punctuation, multiple white spaces, numeric values, stop words and words with size smaller than 4 were removed. Finally, we enriched the corpus by constructing bi-grams and tri-grams.

We used a Latent Dirichlet Allocation (LDA) model [11] to infer the topics present in our research domain. The tuning of the parameters was done through experimentation and qualitative interpretation of the results achieved. Additionally, the coherence score curve was also used as a reference for parameter tuning and the choice of parameters, which are described in Table 1.

2.6 Abstract embeddings

The embeddings of the abstracts was done using the Doc2Vec algorithm [12] and the hyperparameters are defined in Table 1. This allowed the representation of the corpus in a 25 dimension space and was further reduced using a U-map [13] to allow the visualization of the output in a 2-dimensional space.

Model	Hyperparameter	Value
LDA	Num Topics	8
	Chunk Size	2000
	Passes	20
	Alpha	0.1
	ETA	auto
Doc2Vec	Size	25
	Iterations	100
	Min count	10

Table 1: Hyperparameters used.

2.7 Software Implementation

The analysis and modelling was developed using the Python programming language, along with the Scikit-Learn [14], Gensim [15], Umap-Learn [13] and Networkx [16] libraries. The final network analysis

and visualization was done with Gephi [17]. All functions, algorithms, analyses and results are provided in the GitHub repository of the project.

3 Results & Discussion

The popularity of research in data generation has grown significantly in the past 5 years, as shown in Figure 3. Despite the significant amount of uncited publications, out of the ones published in 2020, 39% have already been used in other works. Although most of the research developed before 2016 was used in other works, the amount of cited research increased significantly after that period.

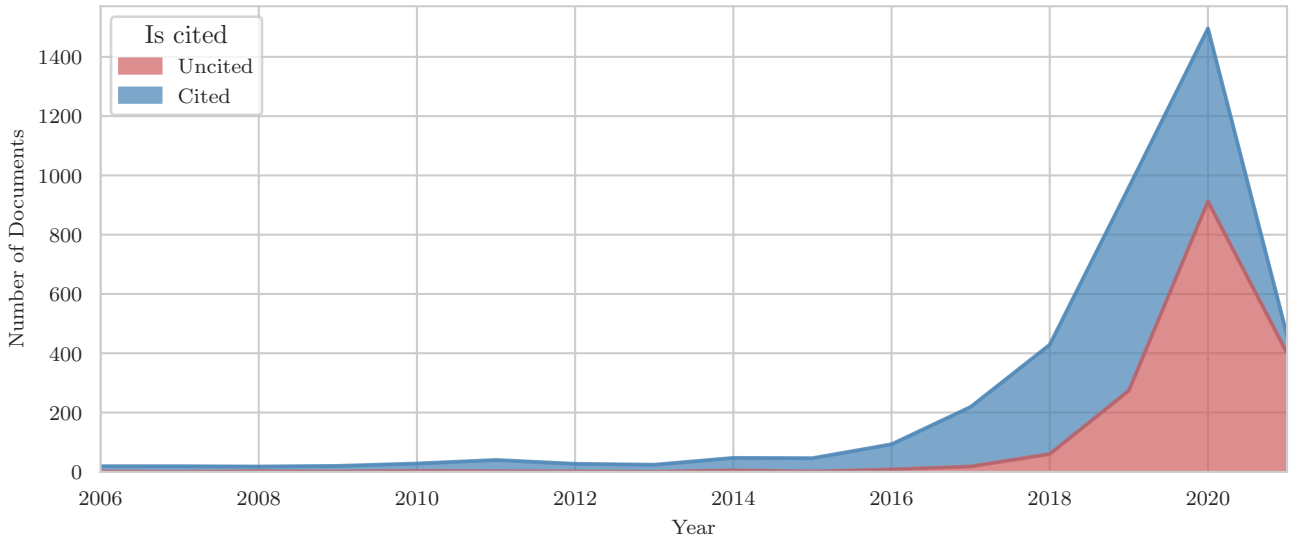

Figure 3: Annual number of publications containing the keyword "data augmentation".

3.1 Journal and Conference Analysis

The initial exploration of the bibliometric data allows us to assess which journals focused in data augmentation more intensely over the past years, as shown in Figure 3. Most of the top journals belong to technical fields, predominantly from Statistics, Remote Sensing, Medical Imaging and other domains of applications such as agriculture. In addition, all these journals have a high impact in their respective fields (based on Scimago Journal & Country Rankings).

Source title	Publications	Citations	Average
Journal of the American Statistical Association	11	538	48.91
IEEE Geoscience and Remote Sensing Letters	19	552	29.05
Neurocomputing	35	808	23.09
Expert Systems with Applications	14	283	20.21
Medical Image Analysis	15	288	19.20
Neural Networks	10	190	19.00
Journal of Computational and Graphical Statistics	23	433	18.83
Computers and Electronics in Agriculture	15	219	14.60
Biometrics	13	163	12.54
IEEE Transactions on Medical Imaging	10	123	12.30

Table 3: Top journals focusing on data augmentation techniques, sorted by citations per document.

Citation-wise, the publications coming from conference proceedings tend to have a comparable impact in the research community, as shown in Table 5. The most relevant conferences are positioned in the computer science and information management fields. Research developed in other areas of application, such as computer vision, speech recognition, acoustic modelling, natural language processing and signal processing have more activity in the form of conference proceedings publications. Conversely, the domains most frequent in journal publications are not as active on conference proceedings publications.

Source title	Publications	Citations	Average
Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	49	2111	43.08
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	372	14946	40.18
Procedia Computer Science	13	288	22.15
International Conference on Information and Knowledge Management, Proceedings	10	180	18.00
IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops	23	314	13.65
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings	95	1153	12.14
Proceedings - International Symposium on Biomedical Imaging	30	346	11.53
Proceedings of the International Conference on Document Analysis and Recognition, ICDAR	17	158	9.29
Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR	13	113	8.69
2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings	12	84	7.00

Table 5: Top conferences focusing on data augmentation techniques, sorted by citations per document.

The papers with the highest citation count are listed in Table 7. We found that much of the research focused on improving deep learning classification, segmentation or object detection without a focus on a particular domain of application. Other papers centered in the application of data augmentation methods for biomedical image classification and segmentation, sound and speech recognition and remote sensing.

Authors	Title	Year	Cited by
Ronneberger O., Fischer P., Brox T.	U-net: Convolutional networks for biomedical image segmentation	2015	13597
Chatfield K., Simonyan K., Vedaldi A., Zisserman A.	Return of the devil in the details: Delving deep into convolutional nets	2014	1885
Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S.	X-Vectors: Robust DNN Embeddings for Speaker Recognition	2018	636
Shorten C., Khoshgoftaar T.M.	A survey on Image Data Augmentation for Deep Learning	2019	590
Salamon J., Bello J.P.	Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification	2017	505
Eitel A., Springenberg J.T., Spinello L., Riedmiller M., Burgard W.	Multimodal deep learning for robust RGB-D object recognition	2015	352
Ding J., Chen B., Liu H., Huang M.	Convolutional Neural Network with Data Augmentation for SAR Target Recognition	2016	319
Wong S.C., Gatt A., Stamatescu V., McDonnell M.D.	Understanding Data Augmentation for Classification: When to Warp?	2016	302
Frid-Adar M., Diamant I., Klang E., Amitai M., Goldberger J., Greenspan H.	GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification	2018	296
Bilen H., Vedaldi A.	Weakly Supervised Deep Detection Networks	2016	287

Table 7: Top papers using data augmentation techniques, sorted by citation count.

3.2 Keyword Analysis

The keyword network shown in Figure 4 revealed 8 main communities of keywords, and 13 other small communities. The different communities are distinguished by the type of algorithms used and/or the domain of application. The main distinctive factor for the larger communities are the types of generative models used, while the smaller communities are distinguished according to the domain of application. The most significant findings we found from this analysis are:

1. The community marked with pink-colored nodes is characterized by the usage of neural network-based data augmentation methods in convolutional neural networks. The keyword "deep learning" is positioned as a central node (although not labelled in the figure to maintain readability). Other relevant keywords are related to machine/deep learning frameworks, deep learning classifiers and data augmentation algorithms, such as "tensorflow", "keras", "convolutional neural network" and "generative adversarial networks". Domain specific keywords are also present:

- Medical keywords located in this community cover a variety of applications. Relevant sub communities are ["hand writing", "parkinson's disease (pd)", "transfer learning"], ["breast cancer", "computer-aided detection"], ["melanoma", "skin cancer", "image processing", "googlenet"], ["chest x-ray", "computer-aided diagnosis", "tuberculosis", "segmentation"] and ["brain", "mri", "multiple sclerosis"].

- Remote sensing keywords are typically related to classification and object detection tasks. Relevant sub communities are ["object detection", "aerial image", "drone", "generative adversarial network", "semantic segmentation"], ["attributed scattering center (asc)", "synthetic aperture radar (sar)", "convolutional neural network (cnn)"], ["remote sensing", "road extraction", "transfer learning", "generative adversarial network"]. Keywords such as "hyperspectral imaging" and "weather classification" are also scattered around the community.

- Facial recognition research is also represented in few sub communities: ["micro expression recognition", "small training data", "convolutional neural network (cnn)", "local binary pattern-three orthogonal planes (lbp-top)"] and ["training data augmentation", "sequence-to-sequence speech synthesis", "sequence-to-sequence speech recognition"].

- Fault detection studies also used data augmentation to deal with imbalanced datasets: ["fault diagnosis", "imbalanced data", "gan"]

- Data augmentation was also associated to regularization methods and feature extraction tasks, based on the presence of the sub communities ["overfitting", "dropout" and "cnn"] and ["feature extraction", "cnn", "svm"].

2. The community marked with blue-colored nodes is characterized by the usage of Markov Chain-based algorithms. The keywords "markov chain", "data augmentation algorithm" and "monte carlo" appear as central nodes. No application-specific sub-community was found.

3. The community marked with green-colored nodes is characterized by the usage of Markov Chain and Bayesian-based algorithms. The keywords "bayesian inference", "markov chain monte carlo", "mcmc", "bayesian analysis", "missing data" and "em algorithm" (expectation maximization algorithm). Application-specific keywords may be found sparsely distributed across the community, all of them related to biological applications. Specifically, the sub community ["ecological health", "stressor-response", "biological monitoring", "bayesian methods"] and the keyword "camera trapping" were found in this community.

4. The community marked with orange-colored nodes is characterized by keywords specific to big data and data warehousing applications. The network is composed of the keywords "big data", "data lake", "olap", "map reduce", "cmm", "data warehouse", "augmentation" and "dm".

5. The remaining communities consist mostly of data augmentation methods applied to specific domains. Specifically, the usage of temporal-dynamic neural network architectures with "eeg (electroencephalogram)", music information retrieval applications (e.g., "chord recognition"), speech/ speaker recognition and embedding, time series forecasting of diabetes and natural language processing and text classification.

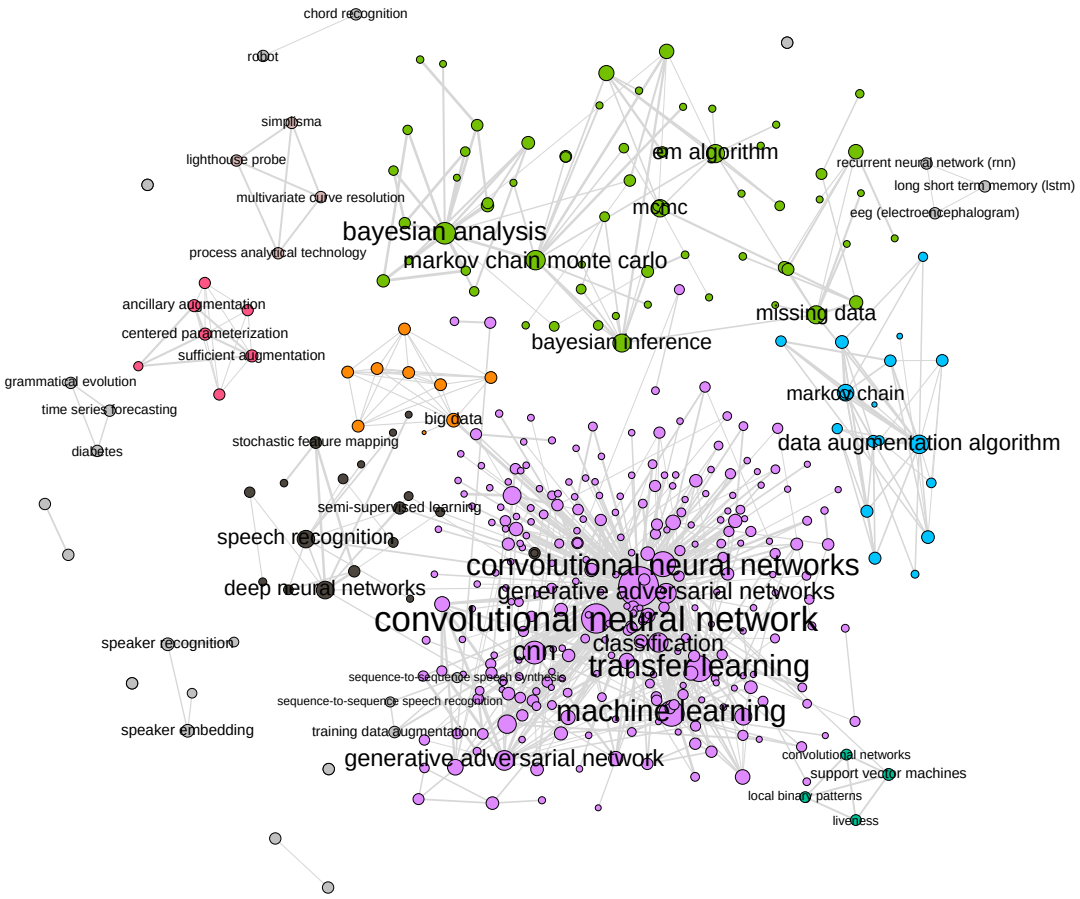

Figure 4: Keyword network.

3.3 Topic Analysis

The LDA topic extraction resulted in 8 different topics, whose distribution of topics is shown in Figure 5. The main topics within which most articles were included is topic 5, which is defined by the main theoretical keywords related to image data augmentation. Rather, the secondary topic is more useful for this analysis. It is found based on the topic likelihood of each document, excluding the dominant topic. Documents belonging to the same group across primary, secondary and/or tertiary topics had a likelihood of zero of belonging to any other topic.

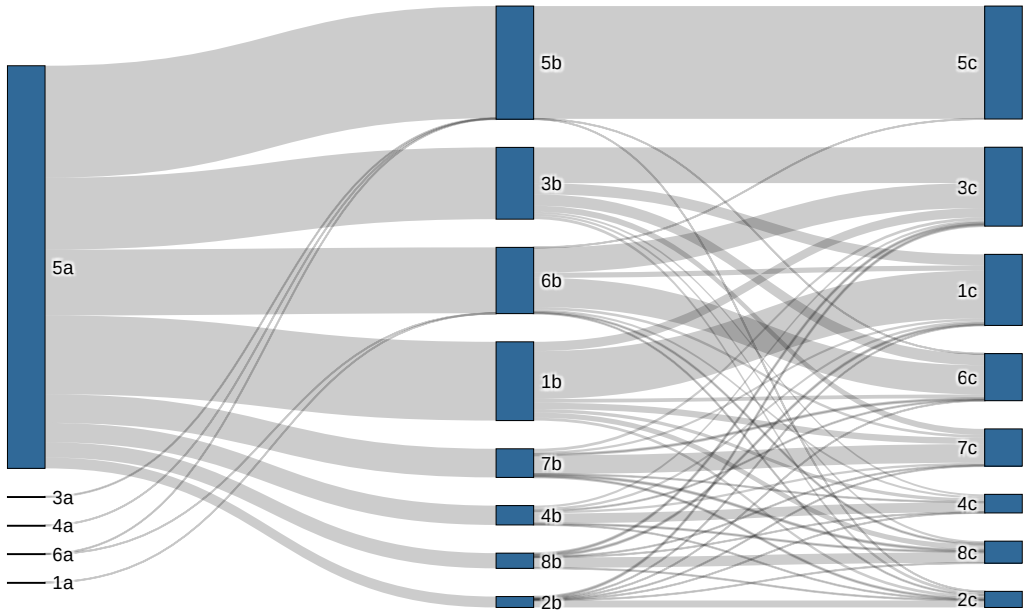

Figure 5: Distribution of documents over the different topics found. The left column represent the primary topics, the middle column represents the secondary topics and the right columns represents the tertiary topics.

The topics found in the bibliometric data are shown in Table 9. A few topics seem to overlap each other, although they are generally distinguishable. The primary domains of application of data augmentation methods differ for each different topic identified:

1. Documents in Topic 1 frequently use the word "yolov", which refers to the YOLOvX family of deep learning object detection models [18], where X refers to the version of the model used (the most recent version is 5). Another relevant keyword is "style_transfer", which refers to a specific technique of data augmentation.

   This topic has two primary domains of application. The keywords "pest" and "coffe" refer to data augmentation on agriculture research. The keywords "biomed", "histolog" and "nodul" refer to biomedical applications such as pulmonary nodule detection and histology image classification. Within these topics, a few domain-specific data augmentation algorithms were proposed. For example, in [19] the authors propose a style-transfer data augmentation method for histology image classification.

2. Documents in Topic 2 are primarily associated to the study of applications that include image data augmentation. The dominant keyword, "hyperspectr_imag", refers to the application of data

augmentation on hyperspectral images, commonly used in remote sensing and medicine. Other classification tasks include license plate detection ("licens_plate"), inpainting ("inpaint"), background subtraction ("illumin_chang") and cloud shadow detection/segmentation ("shadow").

3. Documents in topic 3 refer to the application of data augmentation to deal with censored data (a condition in which the value of an observation is only partially known) and/or supervised tasks on data structured as graphs. Other domains of application involve chest x-rays classification ("cxr"), epidemiology ("risk_factor") and few audio/music information retrieval ("sourc_separ") articles.

4. Documents in topic 4 refer to the application of data augmentation methods on object detection tasks. Specifically fire and smoke, pedestrians and crowd counting. Other applications within this topic are focused on speech recognition and angiography segmentation/classification.

5. Documents in topic 5 are focused on image segmentation and classification methods where data augmentation algorithms are involved. It includes common keywords present in a large set of articles. These articles are mainly focused on the development of different convolutional neural network architectures ("cnn") and neural network-based data augmentation methods.

6. Documents in topic 6 are focused on Bayesian-based algorithms and Markov Chain algorithms. This topic includes data augmentation on regression tasks and misclassification detection.

7. Documents in topic 7 covers the application of data augmentation into various domains. Specifically, music information retrieval ("music"), fish/marine organisms recognition, gender bias, speech recognition, random erasing

8. Documents in topic 8 contains remote sensing and biomedicine as the primary research domains. The keywords "drone" and "aircraft" refer to the sources of data collected for remote sensing work, whereas "pneumonia" and "chest_rai_imag" refers to biomedicine research topics/image data.

Topic	Representative Paper	Papers	Words
1	GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification	440	yolov, pest, style_transfer, coffe, thermal, biomed, scene_text, histolog, nodul, visibl
2	CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training	61	hyperspectr_imag, licens_plate, command, inpaint, illumin_chang, upper, restor, ann, foreign, shadow
3	A survey on Image Data Augmentation for Deep Learning	401	censor, markov_chain, node, team, tree, cxr, risk_factor, mass, largest, sourc_separ
4	Return of the devil in the details: Delving deep into convolutional nets	108	smoke, pedestrian, transcrib, crowd, children_speech, intent, adult, auxiliari_variabl, speech, angiographi
5	U-net: Convolutional networks for biomedical image segmentation	632	imag, detect, gener, dataset, classif, sampl, network, cnn, featur, augment
6	Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification	370	tea, multivari, markov_chain_mont_carlo, bayesian, regress, misclassif, procedur, famili, illustr, mcmc
7	Weakly Supervised Deep Detection Networks	160	music, fish, marin, gender, vocal, random_eras, low_qualiti, crowd, prune, bengali
8	An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare	85	drone, gait, aircraft, gestur_recognit, pneumonia, chest_rai_imag, covid, walk, onset, hidden_layer

Table 9: Description of the main topics found in the literature.

The per-year popularity of the different topics is shown in Figure 6. Since 2015, topic 5 gained more research momentum, whereas topic 6 lost much of its relative popularity within the field. In the past 5 years topics 8 and 3 have become steady research streams while topic 1 saw a significant growth in popularity.

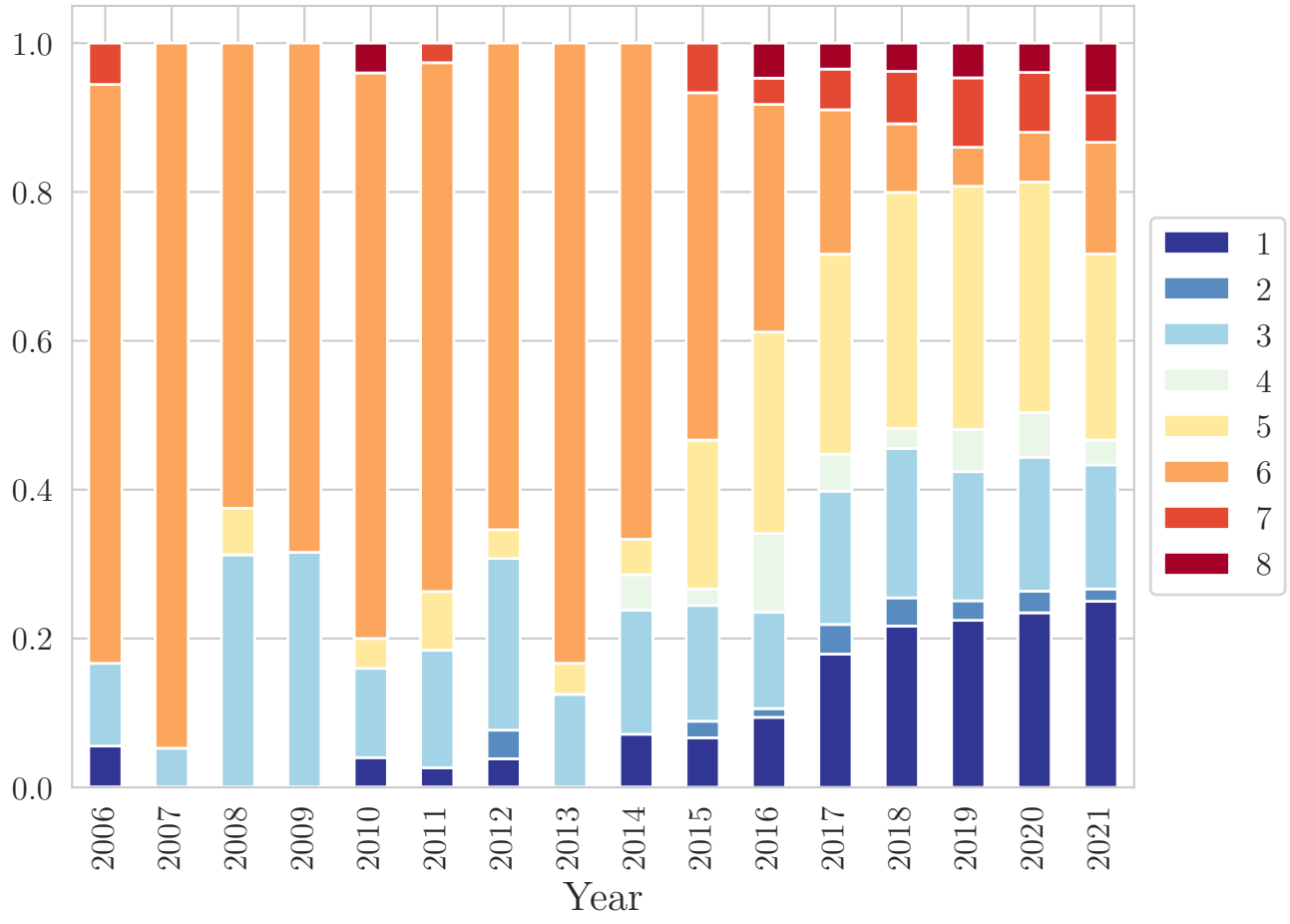

Figure 6: Topic frequency per year.

3.4 Research Gap Discussion

Data augmentation mechanisms are often used as regularization methods for deep learning classifiers. The study of data augmentation mechanisms in ensembles of simple classifiers have achieved state-of-the-art performance not only 10 years ago [4, 5], but also when compared to modern deep learning architectures [6, 7, 20]. However, the implementation of different data augmentation methods shows a promising path to improve the performance of simple classifiers (and/or recent ensemble architectures) and requires further research.

A research application that was not frequently found in the literature was small dataset augmentation. This is particularly useful for any complex problem when the amount of labeled data available to use as training data is scarce, which limits the usage of classification algorithms and especially deep learning algorithms. In this context, techniques such as Active Learning can be used to annotate a small amount of data, while maximizing the classification performance [21]. However, classifiers may not be capable of generalizing with small training datasets and the ability to reproduce and augment the labelled data available can further reduce annotation cost and allow the usage of data intensive classifiers.

Another limitation found in the literature relates to the problem of initialization on network-based data augmentation methods. The same data augmentation algorithm trained with different initialization

settings (different random seed or training subset) may lead to different model parameters and quality of the trained classifier.

The rapid development of data augmentation algorithms raises additional open questions on how the data used and store for model training. Specifically, the lower data storage and processing power available to the general public (*i.e.*, organizations and individuals) is a limitation for producing state-of-the-art classifiers. Another problem arises from data privacy concerns, since the usage of user data to train machine learning models typically involve the storage of such information. However, if data augmentation algorithms were continuously updated and capable of producing reliable data on an as-needed basis, not only would storage requirements decrease, but it would also become possible to work with fully artificial data, without the need to store as much data. This would also facilitate the sharing of datasets (in the form of an algorithm) without compromising sensitive data.

3.5 Study Limitation Discussion

The design of the search query was done based on a single keyword. Although this reduces possible bias, it may have been too broad and didn't include some significant research papers from related techniques such as oversampling. The design of the LDA analysis involved a significant amount of time spent on parameter tuning. Although, due to the subjectivity of the subject, other configurations may be tested in order to optimize the results.

4 Conclusion

Depending on the domain of application, data augmentation research differs in the format of publication. On the one hand, domains like Statistics, Remote Sensing and Medical Imaging seem more active on journal publications, typically in journals with high impact factor. On the other hand, research developed in the domains of computer vision, speech recognition, acoustic modelling, natural language processing and signal processing seem to attribute higher importance to conference papers. Many of the influential papers we found were focused on deep learning methods for classification, segmentation, sound and speech recognition and remote sensing.

We analysed the different communities of keywords formed using document keywords, as well as topic analysis using a LDA analysis over the document's abstracts. We found various distinctive areas of research, both regarding the data augmentation methods used and the domain of application. We found that in recent years research on augmentation methods using Bayesian-based algorithms, as well as Markov Chain algorithms reduced its popularity, whereas data augmentation methods based on neural networks and deep learning classifiers have increased its popularity.

Data augmentation is most commonly applied/studied in the realm of computer vision for tasks like image classification, segmentation, object detection inpainting and background subtraction tasks, even though it may be applied to many other data structures. It is frequently used in studies within the domains of biomedicine, agriculture, speech recognition, acoustic modelling, remote sensing and computational creativity. It is also used alongside other data preprocessing techniques, such as feature extraction and dimensionality reduction.

Although data augmentation is a vibrant area of research, there are still significant gaps to be addressed. Data augmentation methods are increasingly used as regularization methods for deep learning. Although, recent research shows that the same can be done for simpler classifier configurations in order to achieve a classification performance comparable to that of state-of-the-art deep learning, which require further confirmation, as well as the development of less computational intensive data augmentation methods. Other less popular topics, such as small data augmentation, appear to have a relevant practical importance and require further research. In addition, other limitations of data augmentation algorithms should be addressed. One problem commonly found in the literature is the impact the weights initialization and training set used have in the quality of the trained algorithm. In the future, using data augmentation methods as a source of artificial datasets can address a variety of concerns, such as data privacy, sharing and storage. Finally, exploring data augmentation algorithms to complement or replace techniques such as Active Learning may reduce the cost of data collection, although it is yet to be explored.

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, jul 2021.

[2] L. Hu, C. Robinson, and B. Dilkina, "Model Generalization in Deep Learning Applications for Land Cover Mapping," aug 2020.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, pp. 1–48, dec 2019.

[4] U. Meier, D. C. Cireşan, L. M. Gambardella, and J. Schmidhuber, "Better digit recognition with a committee of simple neural nets," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 1250–1254, 2011.

[5] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Handwritten Digit Recognition with a Committee of Deep Neural Nets on GPUs," mar 2011.

[6] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP Architecture for Vision," may 2021.

[7] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "ResMLP: Feedforward networks for image classification with data-efficient training," may 2021.

[8] L. Melas-Kyriazi, "Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet," 2021.

[9] N. Paskin, "Toward unique identifiers," *Proceedings of the IEEE*, vol. 87, pp. 1208–1227, July 1999.

[10] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.

[11] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[12] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *31st International Conference on Machine Learning, ICML 2014*, vol. 4, pp. 2931–2939, International Machine Learning Society (IMLS), may 2014.

[13] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[15] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. http://is.muni.cz/publication/884893/en.

[16] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11–15, 2008.

[17] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, 2009.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 779–788, IEEE Computer Society, jun 2016.

[19] P. A. Cicalese, A. Mobiny, P. Yuan, J. Becker, C. Mohan, and H. V. Nguyen, "StyPath: Style-Transfer Data Augmentation for Robust Histology Image Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12265 LNCS, pp. 351–361, Springer Science and Business Media Deutschland GmbH, oct 2020.

[20] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay Attention to MLPs," may 2021.

[21] T. Su, S. Zhang, and T. Liu, "Multi-spectral image classification based on an object-based active learning approach," *Remote Sensing*, vol. 12, p. 504, feb 2020.