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

1 Introduction

Introduction goes here.

2 Theory

Jürgen Schmidhuber's group shows that a simple MLP architecture can achieve state-of-the-art performance on computer vision benchmarks given strong enough data augmentation [1,2].

[1] Better digit recognition with a committee of simple Neural Nets. Meier, Cires, Gambardella and Schmidhuber 2011 [PDF]

[2] Handwritten Digit Recognition with a Committee of DeepNeural Nets on GPUs. Ciresan, Meier, Gambardella and Schmidhuber 2011 [PDF]

3 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 3.1. The data and text preprocessing is described in Subsection 3.2. The exploratory data analysis described in Subsection 3.3 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 3.4) and study the different communities of keywords found in the network. The topic modelling and parameter tunning is described in Subsection 3.5. The abstract embeddings procedure is described in Subsection 3.6.

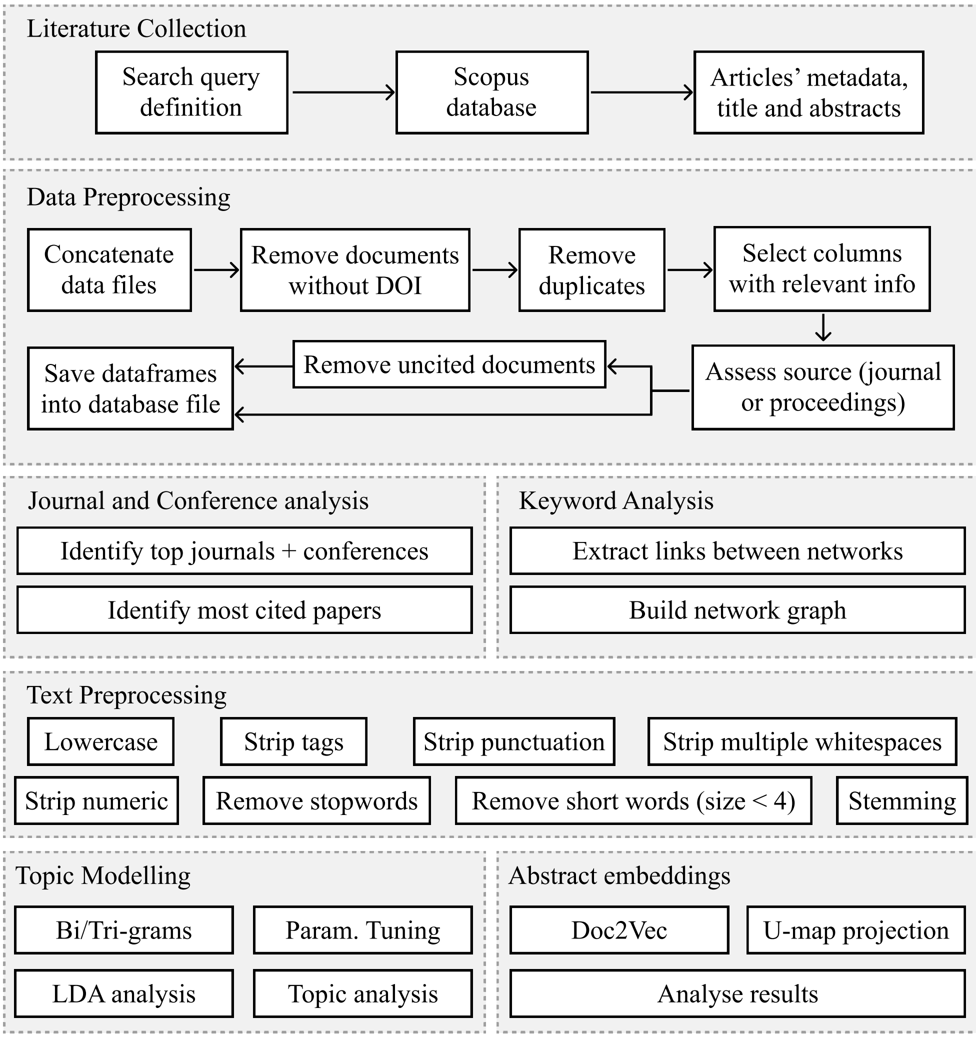

Figure 1: Diagram of the proposed literature analysis approach.

3.1 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 use the keyword "data augmentation" in order to ensure an unbiased analysis. 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 using solely the Scopus database was found to be sufficient. 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 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.

3.2 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 [1], 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 3.3. 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.

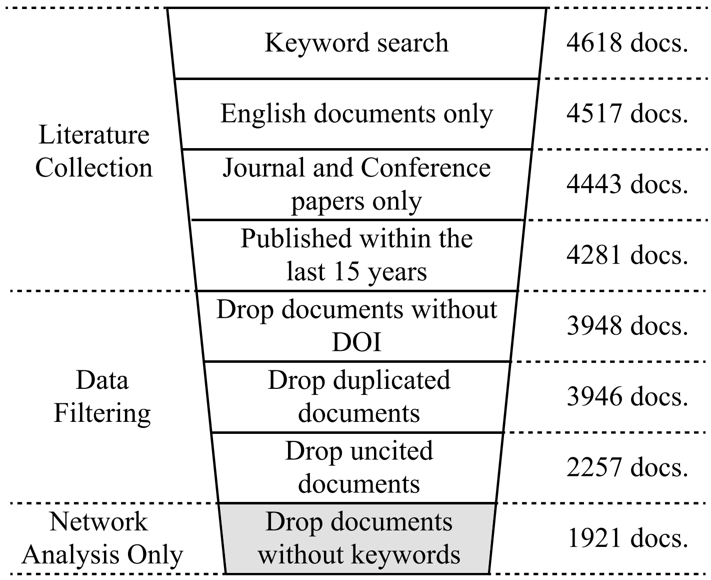

Figure 2: Data filtering pipeline.

3.3 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.

3.4 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.

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 algotithm proposed in [2]. The results of the analysis and community detection were ported to Gephi to produce the final visualizations.

3.5 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 [3] 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 area under the coherence score curve shown in Figure 3 was also used as a reference for parameter tuning and the choice of parameters is described in Table 1.

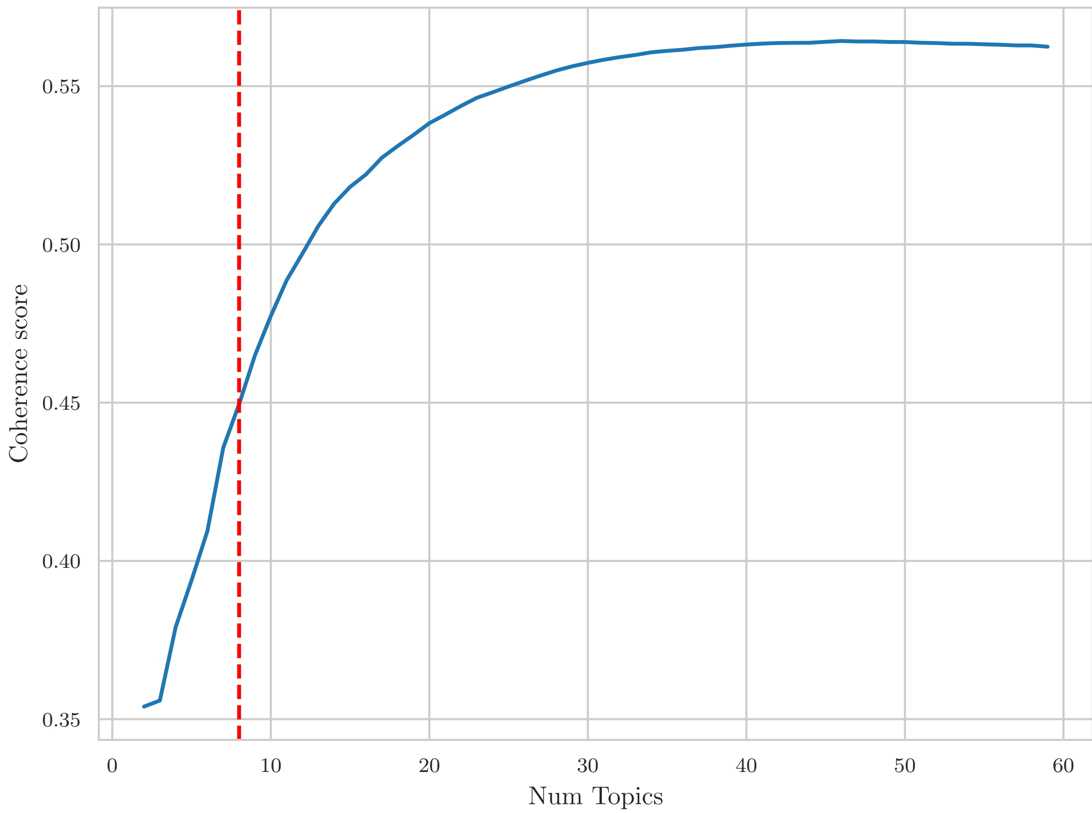

Figure 3: Analysis of the average coherence scores on an optimized LDA model with varying number of topics. The red dotted line represents the number of topics chosen for the final model. The coherence score curve using the average coherence of 5 different initializations over each number of topics.

3.6 Abstract embeddings

The embeddings of the abstracts was done using the Doc2Vec algorithm [4] 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 [5] 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.

3.7 Software Implementation

The analysis and modelling was developed using the Python programming language, along with the Scikit-Learn [6], Gensim [7], Umap-Learn [5] and Networkx [8] libraries. The final network analysis and visualization was done with Gephi [9]. All functions, algorithms, analyses and results are provided in the GitHub repository of the project.

4 Results

The popularity of research in data generation has grown significantly in the past 5 years, as shown in Figure 4. Despite the significant amount of uncited publications, out of the ones published in 2021 up to this date, **XX%** have already been used in other works.

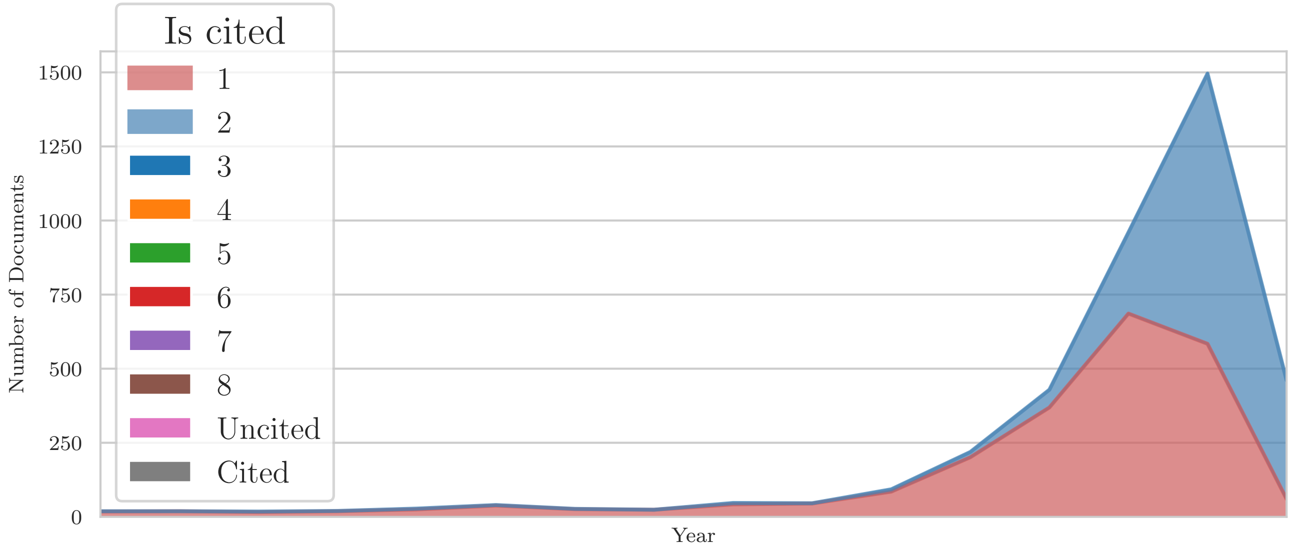

Figure 4: Annual number of publications containing the keyword "data generation".

4.1 Terms Frequency

4.2 Topics Discovered

LDA analysis goes here

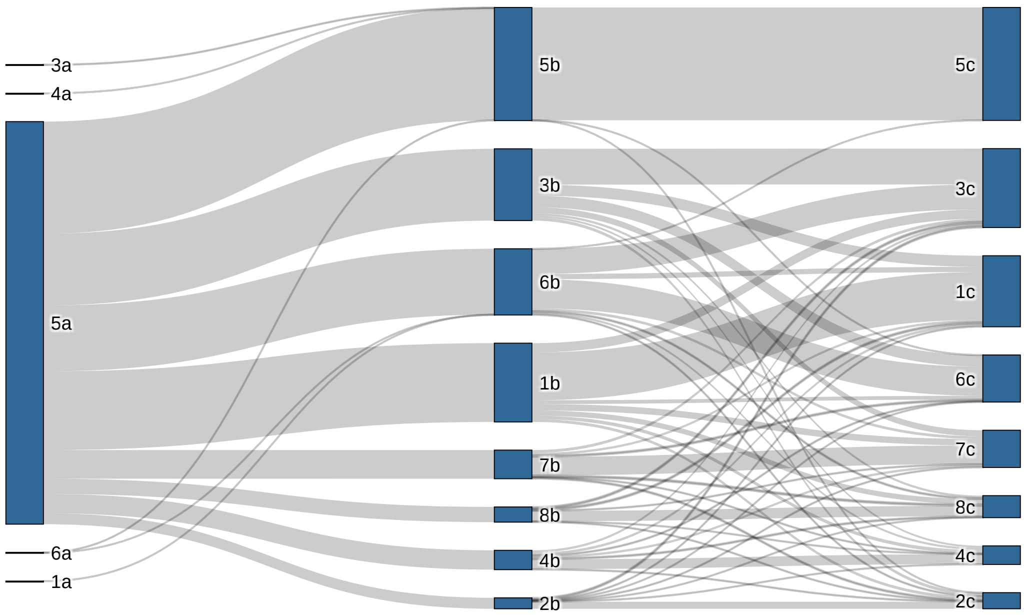

Figure 5: Distribution of documents over the different topics found. The left column represent the primary topics, the middle column represents the secondary topics and the right columns represents the tertiary topics.

4.2.1 Main Journals

4.2.2 Main Conference Proceedings

4.3 Author Co-occurrence Analysis

Not sure whether to keep this one.

4.4 Title and Abstract Text Occurrence Analysis

This can be done by text occurrence network visualization or embeddings.

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 3: Description of the main topics found in the literature.

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 5: Top journals focusing on data augmentation techniques.

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 7: Top conferences focusing on data augmentation techniques.

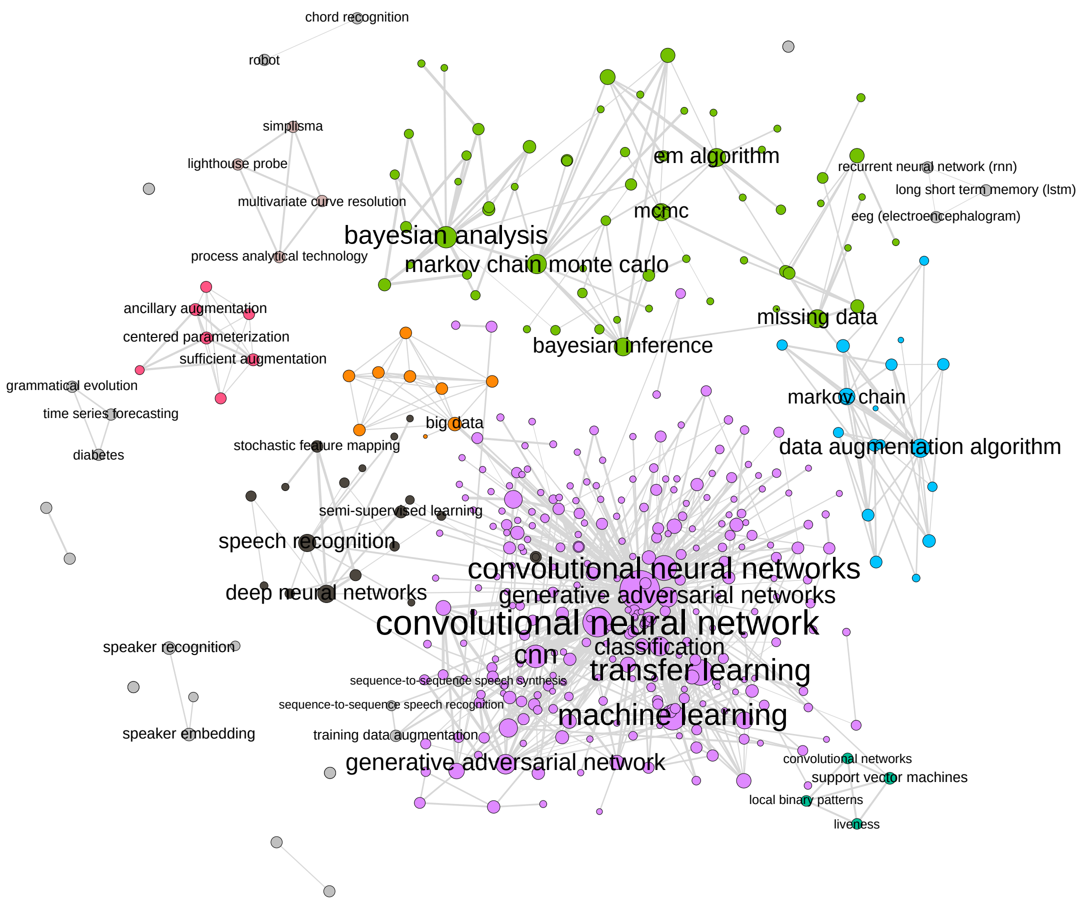

Figure 6: Keyword network.

4.5 Document embeddings

Figure 7: Document embeddings discriminated by LDA topic.

4.6 Most Cited Publications

4.7 Application and Method Analysis

5 Discussion

Discussion goes here.

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 9: Top papers using data augmentation techniques.

5.1 Research Question Discussion

5.2 Research Gap Discussion

5.3 Study Limitation Discussion

6 Conclusions

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.