Document Embedding using piped ELM-GAN Model

Arefeh Yavary
Department of Computer Science
School of Mathematics, Statistics and Computer Science
College of Science
Tehran, Iran
yavary_rf@ut.ac.ir

Hedieh Sajedi
Department of Computer Science
School of Mathematics, Statistics and Computer Science
College of Science
Tehran, Iran
hhsajedi@ut.ac.ir

2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 978-1-6654-2318-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/IMCOM51814.2021.9377413

Abstract— **Document Embedding methods are an impressive task in each machine learning or neural network based natural language processing task. This task is entitled by representation learning and knowledge representation, too. In ultimate the target of this task, each document outputs a representation format of text documents in order to be understandable for machine. Literature reviews in representation learning, shows that document embedding methods for text is weaker in compare with representation of image or signal. Also, in compare to other data like as image or signal, representation of text has more challenges. By this, this paper we suggested a piped process of Generative Adversarial Neural Network and Extreme Learning Machine technique for document embedding. The experimental results show that document embedding using this combination of Generative Adversarial Networks and Extreme learning machines is comparative with other available methods of document embedding.**

Keywords— Document Embedding, Document Modeling, Document Representation, Representation Learning, Knowledge Representation, Extreme Learning Machine, Generative adversarial Network.

I. INTRODUCTION

Document Embedding which entitled in broader concept as representation learning methods, in their ultimate goals comes to account for present an approach to provide understandable representation of document for machine and computer. To make more sense about the complexity of document representation, consider that we have 28 character in English, and different number of sizes of text section like as character, word, sentence, paragraph and document. Each of these mentioned parameters plus considering the different permutations of sections, make documents representation a hard task. Furthermore, for each kind of section in text we need to do a proper document representation. By this document representation is so important in different machine learning and neural network-based model.

In this paper we proposed document mapping with Extreme Learning Machine (ELM) and Generative Adversarial Network (GAN).

Our contributions to this paper are as follows:

• Prepare a group of method for Doc Representation.

• Doc Embedding method for classification and clustering.

• The proposed document embedding method in our research has been evaluated in comparison of most recent and state-of-the-art Document embedding methods.

In the following first we studied recent works in ELM, GAN and also document embedding. Then we proposed our approach for document embedding. After that we demonstrate the results of our experiments and the discussion about the results. At the end we conclude our researches.

II. LITERATURE REVIEW

This paper introduces a method for document embedding using a hybrid method. This hybrid method includes two main model which are GAN and ELM. So, we study the recent researches in review of both of these models. Furthermore, the literature review of this paper contains three parts which are extreme learning machine, generative adversarial neural network and document embedding.

A. Exterem Learning Machine

In the first times, Extreme Learning is provided for classification in a binary class form. Some days after, multi-class form of classification for Extreme Learning method is shown, too. Furthermore, in the pre-knowledge section we introduce the architecture of ELM method in more detail. Also, the recent work in ELM based researches are proposed in the following of this section.

Chen et. al. [1] used kernel-based ELM for Bacterial Foraging Optimization problem. Liu et. al. [2] proposed an optimizer architecture for hybrid swarm-ELM model. Li et. al. [3] suggested ELM for imbalanced learning in problems with sparse cost matrix data.

Chen et. al. [4] considered ELM for feature engineering problem. They used their feature engineering in unsupervised form and for clustering problem.

Wu et. al. [5] combined meta-heuristic algorithms with ELM. They used this architecture is used for monthly pan evaporation prediction model. Yavary et. al. [6] are used ELM for information verification. Hei et. al. [7] utilized ELM for Global solar radiation estimation and climatic variability analysis problem. Their proposed ELM model is a predictive model one.

B. Generavie Adverserial Network

The GAN is one of the interesting and popular neural networks that is very applicable in many machines learning based applications. In the following some of the literature works are provided.

Jiang et. al. [8] used GAN for hyperspectral anomaly detection. Also, they considered discriminative reconstruction constraint in this architecture. Zareapoor et. al. [9] used dual GAN architecture. This architecture used for Perceptual image quality. Glover et. al. [10] proposed an extension of GAN modeling for document representation. Gao et. al. [11] considered GAN model in a wasserstein extension. This

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

neural network is modeled for data augmentation in fault diagnosis problem.

Yang et. al. [12] studied text summarization application with GAN. Their proposed neural network is a multi-constraint task. Chen et. al. [13] utilized GAN with deep architecture and also convolution neural network for study fluid flows. Their Neural network architecture is a data-driven one and studied the model in none-linear form.

Souibgui et. al. [14] used a Conditional Generative Adversarial Network for Document Enhancement. This method could be used for image data too.

C. *Document Embedding*

As we discussed later, the natural aim of document embedding is to represent document such a way that it could be understandable for computer. In this area some historical bag of words models or n-gram are one model, and on the other classic method which is more modern, is the topic models. The most up to date methods are representation which based on vectors. So, in literature review section we focus on vector-based document embedding.

Cohan et. al. [15] suggested a Document-level mapping learning method for citation-informed transformers. This platform entitled by Specter. Brochier et. al. [16] implemented a document embedding model for using inductive document network. This network considered an attention model based on words. Aubaid et. al. [17] proposed a rule-based embedding in order to be used for text classification.

Yao et. al. [18] suggested an embedding network which based on hierarchical labeling. This embedding used for financial document and sentiment analysis.

Dai et. al. [19] used document embedding in order to be used for ad-hoc search. This embedding utilized term weighting and also context in its modeling. Kesiraju et. al. [20] utilized GAN for document embedding and considering uncertainly. This method is shows great results in perplexity measure. Zhang et. al. [21] proposed state-of-the-art method for document representation. We selected this for our evaluation and benchmark in experiments. The embedding method is mostly based on LSTM and considered topics in its mapping.

III. Suggested Method

In this paper we proposed a family of neural network-based document embedding method using GAN and ELM. So, in this section, first we describe the ELM and GAN neural networks in Pre-Knowledge section. Then in proposed method section the architecture of our neural network model is described.

A. *Pre-Knowledge*

This section contains the explanation of two neural networks: Generative Adversarial Network (GAN) and Extreme Learning Machine (ELM).

Generative Adversarial Network: In this neural nnetwork, two neural networks compete with each other in a game two create best weighting in order to optimize the result of the neural network in a task, application or problem.

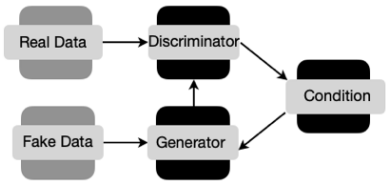

Fig. 1. The artitecture of Generative Adverserial Network.

Extreme Learning Machine: Extreme Learning Machine (ELM) are a kind of feedforward neural network. This network has a lot of usage in several problems like as classification, regression, clustering, sparse approximation, compression and feature learning. This network has a special learning method with mathematical solution which makes this network very impressive. In the Fig. 2. the architecture of the ELM model is illustrated. This architecture is shown in two case of binary and multi-class from for extreme learning machine.

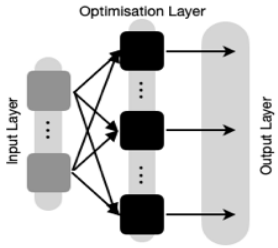

Fig. 2. Structure of the Exterem Learnig Machine in different and basic machine learnig methods.

B. *Proposed Method*

In this section we explain the suggested methods for document embedding. These methods are:

- ELM (ELM-based document embedding)
- ELM Auto-Encoder
- GAN (GAN-based document embedding)
- GAN Auto-Encoder
- ELM-GAN
- ELM-GAN Auto-Encoder

Each of these methods are describe as follows:

ELM as a Document Embedding: In this document mapping we used last layer of ELM as document representation. In Fig. 3. the artitecture of the ELM method for document embedding is illusterate.

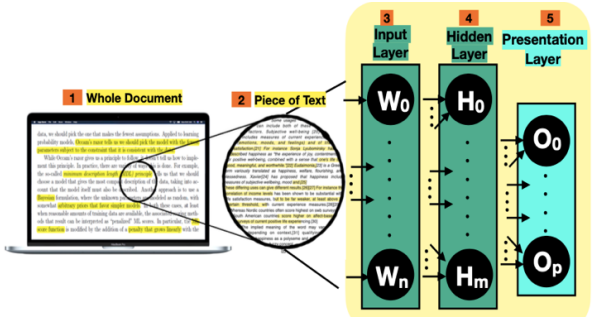

Fig. 3. The artitecture of the ELM method and steps of this process for document embedding.

Auto-Encoder ELM as a Document Embedding: In this modeling, we used Auto-Encoder model of ELM which is consists of two mirrored ELM for our prepose. Then the middle layer is used as the document representation. In the Fig. 4. the artitecture of the ELM AutoEncoder Method is for textual representation learning is shown.

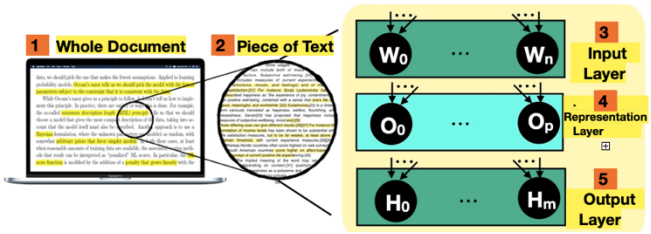

Fig. 4. The artitecture of the ELM AutoEncoder Method and steps of this process for document representation learning.

GAN as a Document Embedding: In this document mapping we used last layer of GAN as document representation.

ELM-GAN as a Document Embedding: This network is a pipelined network which stacked ELM and then GAN network to be used document mapping. We used the last layer as document representation.

ELM-GAN Auto-Encoder: This network is a pipelined network which stacked ELM Auto-Encoder and then GAN Auto-Encoder to be used document mapping. We used the middle layer as document representation.

In the next section we describe the result of experiments of each modelling for document representation.

C. Experimental Results

As we mentioned in the previous section, we proposed six embedding method for document embedding. These methods are: ELM, ELM Auto-Encoder, GAN, GAN Auto-Encoder, ELM-GAN, ELM-GAN Autoencoder. These methods must come to account for evaluation.

We all know that document embedding methods are good if they are good enough in machine learning applications. Between different machine learning applications, Classification as a supervised method and Clustering as an unsupervised method are two of the fundamental and important application of machine learning methods. So, for evaluating document embedding, we selected three clustering method and one classification method. The clustering methods are more because clustering methods are harder than classification. The classification method is Support Vector Machine (SVM) and clustering methods are: K-Means, 1-Nearest Neighbor (1-NN) and N-Cut.

The datasets that we considered for document embedding are three of most popular datasets for text mining, document classification and document clustering. These datasets are: 20 News Group, Routers and Amazon Reviews.

The Table. I up to IV show the result of the experiments. Each Table shows a classification or clustering method. The Table I to IV are K-Means, N-Cut, 1-NN and SVM, respectively.

We used average accuracy of evaluating classification and clustering methods. This is because we want to evaluate one of the best and recent documents embedding methods [20]. The result of this method [20] as entitled by TE-LSTM+SC is reported in a classification task in Table V.

TABLE I. AVERAGE ACCURACY OF K-MEANS

Approach	Data Set		
	20 News Group	*Wiki10+*	*Amazon Reviews*
ELM	0.526	0.434	0.4
ELM Auto-Encoder	0.538	0.434	0.4
GAN	0.558	0.47	0.4
GAN Auto-Encoder	0.565	0.455	0.480
ELM-GAN	0.570	0.460	0.495
ELM-GAN Auto-Encoder	0.589	0.473	0.532

Fig. 5. Experimnetal resutls of different proposed approach for document embedding on three datasets: 20 News Group, Wiki10+ and Amazon Reviews for K-MEANS clustering method.

TABLE II. AVERAGE ACCURACY OF N-CUT

Approach	Data Set		
	20 News Group	*Wiki10+*	*Amazon Reviews*
ELM	0.551	0.420	0.450
ELM AutoEncoder	0.560	0.434	0.465
GAN	0.563	0.440	0.470
GAN AutoEncoder	0.581	0.455	0.489
ELM-GAN	0.595	0.473	0.505
ELM-GAN Auto-Encoder	0.623	0.489	0.532

Fig. 6. Experimnetal resutls of different proposed approach for document embedding on three datasets: 20 News Group, Wiki10+ and Amazon Reviews for N-Cut clustering method.

TABLE III. AVERAGE ACCURACY OF 1-NN

Approach	Data Set		
	20 News Group	*Wiki10+*	*Amazon Reviews*
ELM	0.570	0.454	0.335
ELM Auto-Encoder	0.580	0.460	0.667
GAN	0.598	0.495	0.375
GAN Auto-Encoder	0.608	0.500	3.89
ELM-GAN	0.620	0.500	4.08
ELM-GAN Auto-Encoder	0.623	0.510	0.416

Fig. 7. Experimnetal resutls of different proposed approach for document embedding on three datasets: 20 News Group, Wiki10+ and Amazon Reviews for 1-NN clustering method.

TABLE IV. AVERAGE ACCURACY SVM

Approach	Data Set		
	20 News Group	*Wiki10+*	*Amazon Reviews*
ELM	0.545	0.378	0.389
ELM Auto-Encoder	0.640	0.453	0.450
GAN	0.676	0.462	0.489
GAN Auto-Encoder	0.700	0.498	0.550
ELM-GAN	**0.75**	**0.534**	0.638
ELM-GAN Auto-Encoder	**0.764**	0.510	**0.745**

Fig. 8. Experimnetal resutls of different proposed approach for document embedding on three datasets: 20 News Group, Wiki10+ and Amazon Reviews for SVM classification method.

TABLE V. AVERAGE ACCURACY OF TE-LSTM+SC

Approach	Data Set		
	20 News Group	*Wiki10+*	*Amazon Reviews*
TE-LSTM+SC	0.75	0.534	0.736

Fig. 9. Experimnetal resutl of the state-of-the-art method, TE-LSTM+SC [20] with respect to different datasets which is our benchmark for evaluation.

Totally we observe that ELM-GAN Autoencoder method has the best result with respect to the other methods. Also, this method passed the state-of-the-art methods [20].

D. Discussion

We proposed a group of GAN-ELM based document embedding models. Then we evaluate them with classification and clustering methods. After that we comprised these results with the result of a recent state-of-the-art document embedding model [20].

As the results in the tables in the previous subsection shows, between our several methods for document embedding, the accuracy of models from low to high is as follows: ELM, ELM Auto-Encoder, GAN, GAN Auto-Encoder, ELM-GAN, ELM-GAN Auto-Encoder. Besides as the table IV and V shows, the ELM-GAN Auto-Encoder method can overpass the state-of-the-art method [20].

IV. Conclusion and Future Works

Document Representation methods have an effective role on the huge number of textual based tasks. Besides as these mappings are the first step of prepossessing in most of the machine learning and neural network-based models, so they could improve experimental results of the most of the textual based applications. In this paper, we suggested and implemented a document embedding method using ELM and GAN neural networks. This approach is non-iterative and fast, with comparable results with other document embedding methods. For evaluating the result of the proposed document embedding method, we utilized this mapping with clustering and classification applications. These experimental results show that our suggested document mapping has comparative results with other proposed document-based embedding methods. Furthermore, this method can improve the result of other neural network-based applications.

For extending and optimizing the proposed mapping method for document representation, three main roadmaps can be considered: 1. Extending the GAN method. 2. Extending the ELM method. 3. Extending GAN and ELM together. These extension of mapping methods could be complex for optimizing. The computational complexity of these extensions from low to high is as follows: 1. Extending GAN and ELM together, 2. Extending the GAN method, 3. Extending the ELM method. The computational complexity of these potential future extension of the proposed mappings arise from the complexity of learning methods of the neural networks that embedded in GAN or ELM. By this a novelty in learning strategy for GAN or ELM neural networks could be very impressive.

References

[1] Chen, Huiling, Qian Zhang, Jie Luo, Yueting Xu, and Xiaoqin Zhang. "An enhanced Bacterial Foraging Optimization and its application for training kernel extreme learning machine." Applied Soft Computing 86 (2020): 105884.

[2] Liu, Zhi-Feng, Ling-Ling Li, Ming-Lang Tseng, and Ming K. Lim. "Prediction short-term photovoltaic power using improved chicken swarm optimizer-Extreme learning machine model." Journal of Cleaner Production 248 (2020): 119272.

[3] Li, Hui, Xi Yang, Yang Li, Li-Ying Hao, and Tian-Lun Zhang. "Evolutionary extreme learning machine with sparse cost matrix for imbalanced learning." ISA transactions 100 (2020): 198-209.

[4] Chen, Jichao, Yijie Zeng, Yue Li, and Guang-Bin Huang. "Unsupervised feature selection based extreme learning machine for clustering." Neurocomputing 386 (2020): 198-207.

[5] Wu, Lifeng, Guomin Huang, Junliang Fan, Xin Ma, Hanmi Zhou, and Wenzhi Zeng. "Hybrid extreme learning machine with meta-heuristic algorithms for monthly pan evaporation prediction." Computers and Electronics in Agriculture 168 (2020): 105115.

[6] Yavary, Arefeh, and Hedieh Sajedi. "Rumor detection on Twitter using extracted patterns from conversational tree." In 2018 4th International Conference on Web Research (ICWR), pp. 78-85. IEEE, 2018.

[7] Hai, Tao, Ahmad Sharafati, Achite Mohammed, Sinan Q. Salih, Ravinesh C. Deo, Nadhir Al-Ansari, and Zaher Mundher Yaseen. "Global solar radiation estimation and climatic variability analysis using extreme learning machine based predictive model." IEEE Access 8 (2020): 12026-12042.

[8] Jiang, Tao, Yunsong Li, Weiying Xie, and Qian Du. "Discriminative reconstruction constrained generative adversarial network for hyperspectral anomaly detection." IEEE Transactions on Geoscience and Remote Sensing(2020).

[9] Zareapoor, Masoumeh, Huiyu Zhou, and Jie Yang. "Perceptual image quality using dual generative adversarial network." Neural computing and applications 32, no. 18 (2020): 14521-14531.

[10] Glover, John. "Modeling documents with Generative Adversarial Networks." arXiv preprint arXiv:1612.09122 (2016).

[11] Gao, Xin, Fang Deng, and Xianghu Yue. "Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty." Neurocomputing396 (2020): 487-494.

[12] Yang, Min, Xintong Wang, Yao Lu, Jianming Lv, Ying Shen, and Chengming Li. "Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint." Information Sciences 521 (2020): 46-61.

[13] Cheng, M., Fangxin Fang, Christopher C. Pain, and I. M. Navon. "Data-driven modelling of nonlinear spatio-temporal fluid flows using a deep convolutional generative adversarial network." Computer Methods in Applied Mechanics and Engineering 365 (2020): 113000.

[14] Souibgui, Mohamed Ali, and Yousri Kessentini. "DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[15] Cohan, Arman, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. "Specter: Document-level representation learning using citation-informed transformers." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270-2282. 2020.

[16] Brochier, Robin, Adrien Guille, and Julien Velcin. "Inductive document network embedding with topic-word attention." In European Conference on Information Retrieval, pp. 326-340. Springer, Cham, 2020.

[17] Aubaid, Asmaa M., and Alok Mishra. "A Rule-Based Approach to Embedding Techniques for Text Document Classification." Applied Sciences 10, no. 11 (2020): 4009.

[18] Yao, Ping, Qinke Peng, and Tian Han. "Hierarchical Label Embedding Networks for Financial Document Sentiment Analysis." In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, pp. 499-504. 2020.

[19] Dai, Zhuyun, and Jamie Callan. "Context-Aware Document Term Weighting for Ad-Hoc Search." In Proceedings of The Web Conference 2020, pp. 1897-1907. 2020.

[20] Kesiraju, Santosh, Oldřich Plchot, Lukáš Burget, and Suryakanth V. Gangashetty. "Learning document embeddings along with their uncertainties." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2319-2332.

[21] Zhang, Wenyue, Yang Li, and Suge Wang. "Learning document representation via topic-enhanced LSTM model." Knowledge-Based Systems 174 (2019): 194-204.