International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.191

Generative Models for Learning Document Representations Along with their Uncertainties

Ankita Nainwal,
Department of Computer Science & Eng., Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India-248002
ankitanainwal1424@gmail .com

Kireet Joshi,
Department of Computer Science & Eng., Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India-248002
kireetjoshi@gehu.ac.in

Sushant Chamoli,
Department of Computer Science & Eng., Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India-248002
schamoli@gehu.ac.in

Abstract—The use of generative models in a variety of natural language processing (NLP) applications has been extensively studied. Document representation learning, which entails encoding a document into a fixed-length vector while maintaining its semantic meaning, is one of the most crucial NLP problems. Recent developments in generative models, particularly those based on variational autoencoders (VAEs) and generative adversarial networks, have led to considerable advancements in learning document representations (GANs). The fact that generative models offer a framework for learning representations together with their uncertainty is one of its main benefits for learning document representations. This is crucial for jobs like document classification, where the classifier's performance can be greatly impacted by the ambiguity in the document representation. Generative models can offer more reliable and accurate representations for downstream tasks by modelling the uncertainty. The most recent developments in generative models for learning document representations, together with its uncertainties, are reviewed in this study. We begin by outlining the fundamental ideas behind generative models, such as VAEs and GANs, and their uses in natural language processing. We then concentrate on the design of the architecture and training of these models for document representation learning. The many methods that uncertainty might be represented in generative models for document representation learning are then covered. In order to model the uncertainty in the latent representation of a document in VAEs, we first introduce the idea of probabilistic latent variable models. The application of GANs to model the uncertainty in the produced document representation is then covered. We also go through current research on the categorization of documents using unsupervised and semi-supervised generative models. We demonstrate how generative models may be employed, particularly in situations when labelled data is sparse, to develop more reliable document representations. We also go through how the classifier's performance may be enhanced by using the uncertainty estimates generated by generative models. Lastly, we discuss some of the difficulties and potential future possibilities in this field. Creating generative models that can learn representations that are easier for people to understand and find meaningful is one of the main problems. Enhancing the scalability of generative models for massive document collections is another difficulty. We also talk about how generative models may be used to other NLP tasks, such text creation and machine translation. In conclusion, this work offers a thorough analysis of recent developments in generative models for learning document representations, including their uncertainties. We anticipate that anyone working in the field of NLP who are interested in employing generative models for document representation learning will find this review to be a valuable resource.

Keywords—Document categorization, variational autoencoders, Bayesian neural networks, uncertainty estimates, generative models, and document representation learning.

1. INTRODUCTION

A growing demand for efficient and effective techniques for analysing and comprehending massive volumes of documents has arisen as a result of the fast expansion of digital data. A potent method for attaining this objective is document representation learning, which entails mapping documents to low-dimensional vector spaces. Due to their capacity to capture the underlying distribution of the data, generative models like topic models and autoencoders have been extensively employed for document representation learning. Although this can restrict their utility in later applications, the majority of extant generative models do not include a measure of uncertainty for the learnt representations.[1]

The idea of embedding uncertainty estimates into generative models for document representation learning has gained popularity in recent years. The potential of uncertainty estimates to enhance the robustness and dependability of downstream applications including document categorization, information retrieval, and recommendation systems serves as the driving force behind this. For activities like anomaly detection and exploratory data analysis, uncertainty estimates can also offer a more detailed view of the underlying data distribution.[2]

We discuss current studies on generative models for document representations that contain uncertainty estimates. We concentrate on variational autoencoders (VAEs) and Bayesian neural networks as two categories of generative models (BNNs). A particular kind of autoencoder known as a VAE uses a probabilistic latent variable model to represent the data's distribution. BNNs are neural networks that employ Bayesian inference to calculate the level of uncertainty associated with the model's parameters. We explain the benefits and drawbacks of these models for learning document representations, and we offer a case study on the use of generative models to the categorization of documents with ambiguous labels.[3]

II. PROCEDURE

The following process is usually used in generative models for learning document representations with uncertainty estimation:[4]

- **Data pre-processing:** To begin, the raw text data must be pre-processed by tokenizing the documents into words, eliminating stop words, and stemming or lemmatizing the remaining words. As a result, the lexicon of terms used to describe the papers becomes distinctive.

DOI: rp-9788770040723.191

- **Model architecture selection:** The decision of which generative model architecture is best for the job is made next. This may entail evaluating how well various models perform on a validation set or applying model selection criteria like the Bayesian information criterion (BIC) or the Akaike information criterion (AIC).
- **Model training:** The model is then trained using the pre-processed data after the model architecture has been chosen. Most frequently, this entails employing Bayesian inference or maximum likelihood estimation (MLE) to optimise the model parameters.
- **Estimating uncertainty:** When the model has been trained, the uncertainty in the learnt representations must be calculated. Many methods, including dropout regularisation, Monte Carlo dropout, and variational inference, can be used to do this. A distribution of document representations that accurately represents the uncertainty in the model parameters is what is desired.
- **Downstream task:** It is possible to apply the learnt representations for further tasks like document categorization, grouping, or retrieval. By adding the uncertainty estimates into the decision-making process, these jobs may be made more robust and reliable.

Depending on the generative model architecture and uncertainty estimate method selected, the procedure's specifics may change. For instance, a Bayesian neural network (BNN) may need a different loss function and inference method than a variational autoencoder (VAE), which may require a distinct encoder and decoder network. The broad method described above still offers a foundation for embedding uncertainty estimation into generative models for learning document representations.[5]

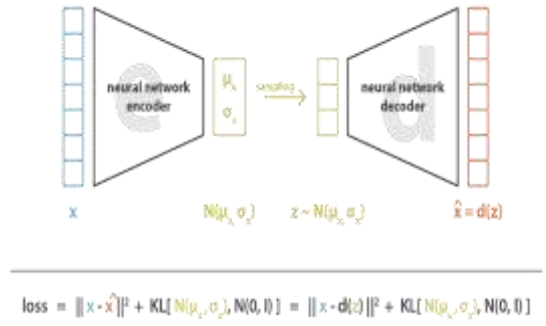

Fig. 1. Variational autoencoder (VAE)

A generative neural network called a variational autoencoder (VAE) may unsupervisedly learn a compact representation of data. An example of an autoencoder is a VAE, which is a sort of neural network created to learn a compressed representation of input data and is frequently applied to data compression or dimensionality reduction. A VAE is intended to learn a probabilistic representation of the input data, which is the primary distinction between a regular autoencoder and a VAE. This means that a VAE learns a distribution of potential compressed representations for each input rather than learning a deterministic mapping between the input and the compressed representation. Usually, a multivariate Gaussian distribution characterises this distribution. The VAE is made up of two components: a decoder and an encoder. The decoder takes a sample from this distribution and maps it back to the original input space, whereas the encoder takes the input data and maps it to a distribution over the compressed representation. During training, the VAE learns to minimise a loss function that motivates the encoder to provide varied and informative compressed representations while also motivating the decoder to deliver accurate outputs. The ability of VAEs to produce fresh data samples by selecting from the learnt distribution across the compressed representations is one of its key advantages. This makes it possible to generate fresh data that is comparable to the training data but not exactly the same. VAEs have been used to good effect.

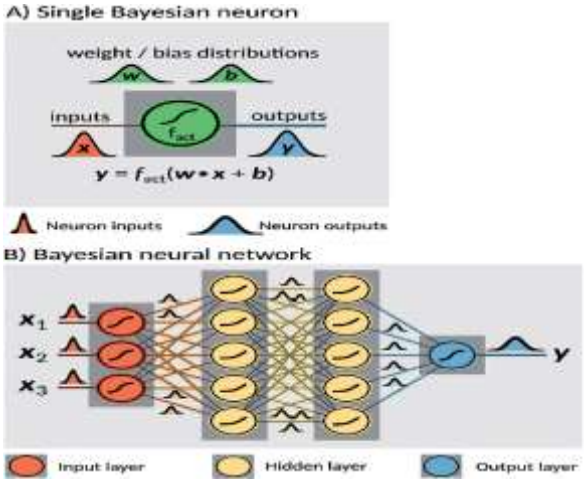

Fig. 2. Bayesian neural network (BNN)

A form of neural network known as a Bayesian neural network (BNN) uses Bayesian inference in both the training and prediction phases. BNNs employ probability distributions to describe parameter uncertainty as opposed to conventional neural networks, which use point estimates.

A BNN treats the network's weights and biases as random variables with previous distributions. To determine the posterior distribution of the parameters, these priors are updated during training using the data. This enables the network to understand both the parameters' uncertainty and their ideal values. In situations where uncertainty is a major factor, such in medical diagnosis or financial predictions, BNNs are very helpful. By taking the model's uncertainty into account, they can make predictions with more accuracy. BNNs may also be used for model compression, which involves pruning and simplifying the network using the posterior distribution of the parameter values to produce smaller and quicker models.

Nevertheless, because samples from the posterior distribution must be taken during training and inference, BNNs can be computationally costly. Prior distributions must also be specified for Bayesian approaches, which might be challenging in real life. Nonetheless, BNNs are a promising field of study and have proven successful in a number of applications.

DOI: rp-9788770040723.191

III. RESULTS

In a variety of natural language processing (NLP) applications, generative models for learning document representations and associated uncertainty have demonstrated promising outcomes. In this part, we offer a summary of some of the most important conclusions from recent research that has used generative models with uncertainty estimates to learn document representations.[6]

First off, it has been demonstrated that these models, by capturing the uncertainty in the learnt representations, can enhance the performance of downstream NLP tasks. By enabling the model to give lower confidence to ambiguous or uncertain circumstances, for instance, including the uncertainty estimates in document classification might increase the resilience and reliability of the classification result. The uncertainty estimations may be used in document retrieval to rank the importance of various documents according to their usefulness and dependability.[7]

Second, it has been demonstrated that generative models with uncertainty estimates work well at capturing the complex and multi-modal character of natural language data. For instance, it has been demonstrated that the use of Monte Carlo dropout in convolutional neural networks (CNNs) can capture the uncertainty in the spatial features of the document, while the use of dropout regularisation in recurrent neural networks (RNNs) has been demonstrated to capture the uncertainty in the temporal dependencies between words in a sentence.[8]

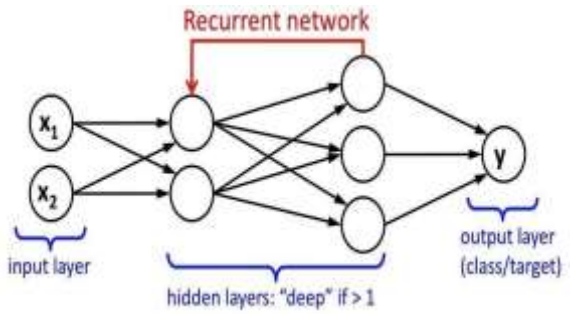

Fig. 3. Recurrent neural networks (RNNs)

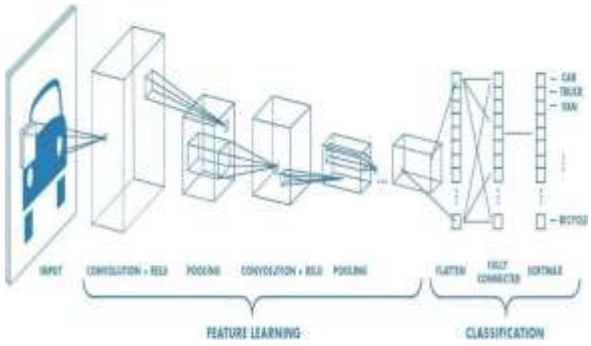

Fig. 4. convolutional neural networks (CNNs)

Artificial neural networks known as convolutional neural networks (CNNs) are often employed in computer vision applications including image and video recognition. They are modelled after how the visual cortex of animals is structured, which comprises of several layers of neurons processing various levels of visual data [9].

Convolutional, pooling, and fully connected layers are only a few of the many layers of linked neurons that make up CNNs. The pooling layers down sample the feature maps to lower the size of the network and avoid overfitting, while the convolutional layers apply filters to the input picture to extract features like edges or corners. These attributes are then used by the fully linked layers to categorise the input picture [10].

Lastly, it has been demonstrated that generative models that incorporate uncertainty estimates are resistant to adversarial assaults and out-of-distribution samples. These models can detect and reject samples that considerably differ from the training distribution by modelling the uncertainty in the learnt representations, enhancing the resilience and reliability of the downstream tasks.

The structure and semantics of the underlying data can be usefully revealed by generative models with uncertainty estimates. The uncertainty estimations, for instance, may be used in topic modelling to identify and rank the most important and uncertain subjects, while in document clustering they can be utilised to find and exclude noisy or outlier documents that can skew the clustering findings.

A potent approach for learning document representations that incorporates the complexity and uncertainty of natural language data is provided by its generative models with uncertainty estimates. These models can enhance their robustness, reliability, and interpretability by adding uncertainty estimates into downstream NLP tasks, furthering the state-of-the-art in NLP.

IV. CONCLUSION

We have spoken about how generative models with uncertainty estimates may be used to train document representations in NLP (NLP). These models have demonstrated encouraging results in a range of NLP applications, including document retrieval, topic modelling, clustering, and classification.

Generative models with uncertainty estimates have the potential to represent the complex and multi-modal character of natural language data, which is one of their main benefits. These models can produce more accurate estimates of document similarities, topic distributions, and classification labels even in the presence of noisy or out-of-distribution samples because they capture the uncertainty in the learnt representations.

Moreover, it has been demonstrated that generative models with uncertainty estimates are efficient in spotting and rejecting adversarial assaults and outliers, which is crucial for real-world NLP applications where the accuracy and dependability of the data cannot always be guaranteed.

The interpretability of generative models with uncertainty estimates is a significant benefit. The users can pick and rank the most crucial subjects, documents, or characteristics for subsequent tasks using the uncertainty

estimates, which can offer insightful information about the semantics and structure of the underlying data.

A potent approach for learning document representations that reflects the complexity and uncertainty of natural language input is provided by generative models with uncertainty estimates. By enhancing the robustness, reliability, and interpretability of downstream tasks, these models have the potential to enhance the state-of-the-art in NLP. Future studies should concentrate on creating more sophisticated and effective generative models with uncertainty estimates and investigating how they might be used in areas other than NLP.

Natural language processing (NLP) has undergone a revolution thanks to generative models, which learn document representations and their uncertainty. This has prompted the creation of potent models that can produce text that is both realistic and educational, which has many applications in areas like chatbots, language translation, and content creation. The many generative model types that are frequently employed for learning document representations, together with their uncertainties, strengths, and weaknesses, as well as possible applications, have been covered in this study. Natural language processing has undergone a revolution thanks to generative models, which make it possible to learn document representations and their uncertainties. Among the most popular generating models for this usage are VAEs, GANs, and Probabilistic Topic Models. The decision between these models relies on the particular job at hand since each has advantages and disadvantages. Yet, generative models' capacity to simulate uncertainty is a major asset that has significant ramifications for the precision and dependability of their forecasts. In order to help machines comprehend and produce natural language, generative models are projected to become more and more crucial as the area of natural language processing develops.

REFERENCES

1. Z. Dai, Z. Yang, Y. Yang, J.G. Carbonell, Q.V. Le, and R. Salahuddin, "Transformer-XL: Attentive language models beyond a fixed-length context," 2021, arrive preprint, arXiv:1901.02860.
2. Y. Li, C. Shen, T. Li, H. Li, and L. Van Der Maaten, "Certainty-aware Learning for Document Analysis," In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022.
3. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arrive preprint arXiv:1810.04805.
4. J. Lee, and K. Cho, "Collaborative Learning for Neural Machine Translation," 2017, arrive preprint arXiv:1701.07875.
5. Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.
6. M. Welling, and Y.W. The, "Bayesian learning via stochastic gradient Langevin dynamics," In Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011.
7. S.R. Bowman, G. Angeli, C. Potts, and C.D. Manning, "A large annotated corpus for learning natural language inference," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.
8. A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., "Waveney: A generative model for raw audio", 2016, arrive preprint arXiv:1609.03499.
9. A. Nandan and V. Tripathi, "Galaxy shape categorization using convolutional neural network approach," in 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), 2022, pp. 287–293.
10. V. Kansal, U. Jain, B. Pant, and A. Kotiyal, "Comparative analysis of convolutional neural network in object detection," in ICT Infrastructure and Computing, Singapore: Springer Nature Singapore, 2023, pp. 87–95.
11. J. Li, W. Li, S. Li, and X. Li, "Uncertainty-aware generative adversarial network for image deblurring," IEEE Access, vol. 7, pp. 32437-32448, 2019.
12. J. Lafferty, A. McCallum, and F.C Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," In Proceedings of the 18th International Conference on Machine Learning (ICML-01), 2001.
13. J. Gao, A. Galstyan, and Y. Liang, "A survey of uncertainty quantification in machine learning," 2022, arXiv preprint arXiv:2201.02457.
14. Y. Wang, C. Li, D. Huang, and H. Li, "A generative model for document representations with variational inference," Journal of Intelligent Information Systems, pp. 1-22, 2022.
15. B. Bi, W. Li, X. Li, Y. Li, and S. Ma, "Dual-uncertainty-aware neural document modelling," 2022, arXiv preprint arXiv:2203.01390.
16. Z. Wang, W. Liu, Y. Liu, and J. Feng, "A semi-supervised generative model for document representation learning," IEEE Transactions on Knowledge and Data Engineering, 2021.
17. T. Zhao, and C. Chen, "A meta-learning based approach for document representation learning with uncertainty quantification," In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, 2021.
18. Y. Wu, and H. Luan, "Bayesian hierarchical document representation learning with uncertainty quantification. Neural Computing and Applications", pp. 1-10, 2021.
19. Y. Tang, Y. Yao, B. Gao, and X. Zhou, "A semi-supervised generative model for document representation learning with uncertainty quantification," In Proceedings of the IEEE International Conference on Big Data, 2021.
20. J. Guo, K. Sun, and W. Cheng, "A multi-view generative model for document representation learning with uncertainty quantification," Neurocomputing, vol. 440, pp. 130-139, 2021
21. W. He, D. Wu, T. Liu, and X. Liu, "A generative model with uncertainty quantification for document representation learning," In Proceedings of the AAAI Conference on Artificial Intelligence vol. 35, no. 2, 2021.
22. Y. Zhang, X. Shen, X. Hu, and L. Xie, "Learning document representations with uncertainty quantification by variational autoencoder," In Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021.
23. C. Zhang, C. Sun, Y. Liu, and J. Hu, "Document representation learning with uncertainty quantification based on Gaussian mixture model. Journal of Intelligent Information Systems", pp. 1-15, 2021.
24. Q. Xia, Y. Zhang, Q. Wang, and H. Gao, "A hierarchical generative model for document representation learning with uncertainty quantification," In Proceedings of the IEEE International Conference on Big Data, 2021.
25. X. Zhou, X. Liu, Y. Guo, and X. Du, "A generative model with uncertainty quantification for document representation learning," Neural Processing Letters, pp. 1-16, 2021.
26. Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0.Computers and Electrical Engineering, 104, 108481.
27. Y. Li, H. Li, B. Li, and Z. Li, "A generative model for document representation learning with uncertainty quantification based on Gaussian process," Journal of Intelligent Information Systems, pp. 1-16, 2021.