Vector Semantics and Embeddings

Raza Hamid
LUMS

Foundations of Generative AI
Agha Ai Raza
10th February 2025

Introduction and Background:

The field of natural language processing (NLP) has traditionally depended on statistical and rule-based techniques for text representation and analysis. Approaches like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) were commonly used but had notable drawbacks. These models often struggled with sparsity, did not effectively capture the contextual meanings of words, and were costly in terms of computational resources for training and implementation. The introduction of distributed word representations marked a significant turning point, providing more efficient and semantically rich methods for encoding language. The paper *Efficient Estimation of Word Representations in Vector Space* by Mikolov et al. [1] was pivotal in this transformation. It presented Word2Vec, a neural network-based algorithm that generates dense, continuous word embeddings. Unlike earlier models, Word2Vec not only preserved the semantic and syntactic relationships between words but also ensured computational efficiency, representing a major leap forward in vector semantics. This research built on previous studies in distributional semantics, particularly Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Neural Probabilistic Language Models, while significantly enhancing scalability and performance. By tackling both efficiency and effectiveness, this work made a substantial contribution to the creation of models capable of probabilistically inferring relationships between words with high accuracy and speed. The main goal of the study was to develop a more computationally efficient approach for learning word representations in high-dimensional vector space. The research aimed to address two critical questions:

- How can word representations be learned efficiently without sacrificing semantic depth?
- Can a neural network-based approach outperform traditional language models in capturing word relationships?

This study marked a significant turning point in natural language processing (NLP) and generative AI. The introduction of word embeddings set the stage for more sophisticated models like GloVe, FastText, and Transformer-based architectures such as BERT and GPT. By addressing important challenges related to efficiency and scalability, Word2Vec emerged as a vital element in deep learning-driven NLP, fueling progress in various applications from machine translation to conversational AI [1,2,3,4].

Methods and Technical Details:

The paper *Efficient Estimation of Word Representations in Vector Space* presents effective methods for learning word embeddings, focusing on the Skip-gram and Continuous Bag of Words (CBOW) models, which are central to the Word2Vec framework [1]. These models create dense vector representations of words based on their usage in context, adhering to the principle of distributional semantics—the notion that words found in similar contexts often share similar meanings. CBOW generates embeddings by predicting a target word from its surrounding context, while Skip-gram does the opposite, predicting context words from a given target word. Both models utilize neural network-based probabilistic methods to enhance word embeddings.

To boost training efficiency, the paper introduces two important optimization techniques. Hierarchical Softmax substitutes traditional softmax normalization with a tree-based method, lowering its computational complexity from O(V) to O(log V), where V represents the vocabulary size [5]. This adjustment allows for the calculation of probabilities over extensive vocabularies. Negative Sampling, an alternative to full softmax, approximates the probability distribution by selecting a small number of negative examples rather than evaluating the entire vocabulary. This approach significantly speeds up training while preserving high-quality embeddings.

The paper emphasizes several key contributions that set Word2Vec apart from earlier methods. Firstly, it achieves notable scalability improvements, making it much more efficient than conventional neural probabilistic language models, which were often computationally intensive [2]. Secondly, it offers two alternative methods—Negative Sampling and Hierarchical Softmax—that remove the necessity for full softmax normalization, which had been a significant bottleneck. The paper *Efficient Estimation of Word Representations in Vector Space* presents effective methods for learning word embeddings, focusing on the Skip-gram and Continuous Bag of Words (CBOW) models, which are central to the Word2Vec framework [1]. These models create dense vector representations of words based on their usage in context, adhering to the principle of distributional semantics—the notion that words found in similar contexts often share similar meanings. CBOW generates embeddings by predicting a target word from its surrounding context, while Skip-gram does the opposite, predicting context words from a given target word. Both models utilize neural network-based probabilistic methods to enhance word embeddings. To boost training efficiency, the paper introduces two important optimization techniques. Hierarchical Softmax substitutes traditional softmax normalization with a tree-based method, lowering its computational complexity from O(V) to O(log V), where V represents the vocabulary size [5]. This adjustment allows for the calculation of

probabilities over extensive vocabularies. Negative Sampling, an alternative to full softmax, approximates the probability distribution by selecting a small number of negative examples rather than evaluating the entire vocabulary. This approach significantly speeds up training while preserving high-quality embeddings. The paper emphasizes several key contributions that set Word2Vec apart from earlier methods. Firstly, it achieves notable scalability improvements, making it much more efficient than conventional neural probabilistic language models, which were often computationally intensive [2]. Secondly, it offers two alternative methods—Negative Sampling and Hierarchical Softmax—that remove the necessity for full softmax normalization, which had been a significant bottleneck.

Results and Analysis:

The paper *Efficient Estimation of Word Representations in Vector Space* outlines a series of experiments aimed at assessing the effectiveness of the proposed word embedding techniques, mainly Skip-gram and Continuous Bag of Words (CBOW). A significant numerical finding is the enhanced performance of these models in capturing word similarities when compared to traditional methods like Latent Semantic Analysis (LSA) and n-gram models. The authors note that the Skip-gram model, in particular, achieves lower perplexity and shows strong semantic and syntactic generalization across various linguistic tasks [1]. Furthermore, the authors perform ablation studies to explore the effects of factors such as vector dimensionality, window size, and negative sampling. The findings indicate that increasing vector dimensionality boosts performance up to a certain point, after which the benefits begin to diminish. Negative sampling greatly improves training efficiency without a substantial loss in accuracy. Additionally, the experiments reveal that larger window sizes are effective in capturing broader contextual relationships, although excessively large sizes can introduce noise [1]. The results strongly reinforce the paper's main assertion that word representations learned through Skip-gram and CBOW are computationally efficient while maintaining meaningful linguistic relationships. The reported enhancements in similarity tasks and analogy completion further confirm the effectiveness of these embeddings. In comparison to earlier methods, these models strike a better balance between efficiency and accuracy.

The paper also points out some limitations. A significant one is the dependence on large-scale data for achieving optimal performance. The models do not perform as well on smaller datasets, indicating that they require extensive training data to effectively understand word relationships. Additionally, while negative sampling helps speed up training, it also introduces a degree of randomness in the optimization process, which can impact the stability of the model. An interesting observation from the paper is that embeddings created with larger window sizes are better at capturing conceptual and topical similarities, while smaller windows tend to emphasize syntactic features. This supports the idea that local contexts influence syntactic roles, whereas broader contexts contribute to semantic coherence.

However, there are some surprising contradictions in certain instances where words with similar meanings do not cluster closely in vector space, suggesting that purely distributional methods may have difficulty with specific linguistic subtleties. In summary, the experimental results validate the effectiveness of the proposed models while also revealing the trade-offs between efficiency and accuracy. These findings open up opportunities for further improvements in word representation techniques, especially in enhancing data efficiency and capturing more profound semantic relationships.

Critical Evaluation:

Efficient Estimation of Word Representations in Vector Space by Mikolov et al. [1] introduces a groundbreaking method for learning word embeddings through the Skip-gram and Continuous Bag of Words architectures. This paper stands out for its computational efficiency and straightforward explanation of the algorithms involved, making it both accessible and reproducible. These qualities have contributed to its widespread use and significant impact in the fields of natural language processing and generative AI. However, the study does have some drawbacks. The dependence on static word embeddings means that the model struggles to capture the different meanings of words in various contexts. This limitation has been addressed by later models like ELMo [7], BERT [8], and GPT [9], which have developed contextualized representations. Additionally, the evaluations were performed on relatively limited datasets, which may restrict the applicability of the results. The paper also does not delve deeply into issues such as hyperparameter sensitivity or the potential for biases in the training data to be perpetuated. Looking to the future, subsequent research could expand on these foundations by incorporating dynamic, context-sensitive representations alongside the efficient architectures proposed by Mikolov et al. [1]. Broadening evaluations to encompass more diverse and multilingual datasets could enhance model robustness, while a more thorough examination of bias mitigation would be essential for ethical applications. Furthermore, additional theoretical investigation into why certain vector operations effectively capture semantic relationships could pave the way for even more advanced and reliable models.

Personal Reflection:

Reading *Efficient Estimation of Word Representations in Vector Space* has greatly improved my understanding of generative AI by offering clear insights into neural language modeling. I learned how the Skip-gram and Continuous Bag of Words models function to capture both semantic and syntactic relationships in language. The paper explained the practical advantages of Hierarchical Softmax and Negative Sampling as effective methods to lower computational complexity and speed up training, making large-scale language modeling achievable [1,5]. These specific techniques have shaped my research focus on creating scalable AI systems. I now have a better understanding of how to utilize efficient embedding methods to develop models that perform well even with low-resource languages and in real-time applications. Furthermore, the discussion on moving from static to context-aware embeddings has motivated me to investigate hybrid approaches that enhance both performance and fairness in language representation. This paper has given me practical insights that will directly influence my future projects aimed at building more robust and effective generative AI systems.

Conclusion:

In summary, *Efficient Estimation of Word Representations in Vector Space* presents a groundbreaking method for natural language processing by creating effective word embedding techniques using the Skip-gram and Continuous Bag of Words models. These approaches, improved by optimization strategies like Hierarchical Softmax and Negative Sampling, tackle significant issues related to scalability and computational efficiency while maintaining semantic and syntactic connections. The contributions of this paper have set the stage for future advancements in generative AI, impacting both research and real-world applications.

References:

[1] Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781.

[2] Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. *A Neural Probabilistic Language Model*. Journal of Machine Learning Research 3:1137–1155.

[3] Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science 41, 6:391–407.

[4] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3:993–1022.

[5] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. *Distributed Representations of Words and Phrases and Their Compositionality*. In *Advances in Neural Information Processing Systems*, 3111–3119.

[6] van der Maaten, L.; and Hinton, G. 2008. *Visualizing Data Using t-SNE*. Journal of Machine Learning Research 9:2579–2605.

[7] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. *Deep Contextualized Word Representations*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, 2227–2237.

[8] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, 4171–4186.

[9] Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. *Improving Language Understanding by Generative Pre-training*. Retrieved from https://openai.com/research/language-unsupervised