International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

On the Dimensionality of Word Embeddings: An Empirical Study of the Accuracy, Latency, and Memory Trade-off in Semantic Retrieval

Abdullah Yousuf
Department of Computer Science

***Abstract* - Dense vector embeddings underpin modern semantic search and retrieval systems, yet the choice of embedding dimensionality is often made by convention rather than measurement. This paper presents a controlled empirical study of how dimensionality affects three quantities that matter in practice: retrieval quality, query latency, and memory footprint. Skip-gram Word2Vec embeddings were trained at five dimensions (25, 50, 100, 200, and 300) on a 2.1-million-word public-domain corpus, then evaluated on two independent benchmarks—the Mikolov word-analogy task and the WordSim-353 similarity-correlation task—while query latency and memory were measured directly. The results show that retrieval quality rises sharply at low dimensions but plateaus by dimension 100, with the similarity benchmark peaking as early as dimension 50 and declining thereafter, whereas latency and memory grow strictly linearly with dimension. Consequently, the highest-dimensional model studied consumed twelve times the memory and over five times the query time of the smallest for no measurable gain in quality. The findings indicate that, for modest corpora, low-to-moderate dimensionality offers the best accuracy-per-resource trade-off, and that increasing dimensionality past the saturation point is not merely wasteful but can degrade similarity performance.**

***Keywords* - word embeddings; dimensionality; semantic retrieval; information retrieval; Word2Vec; efficiency**

I. INTRODUCTION

Semantic retrieval—the task of finding items related in meaning rather than in surface form—has become a foundational component of modern artificial intelligence systems. Search engines, recommendation systems, question-answering pipelines, and the retrieval-augmented generation systems that support large language models all rely on representing words, sentences, or documents as dense vectors in a continuous space. In this space, semantic similarity is approximated by geometric proximity, so that retrieving relevant items reduces to finding nearest neighbours of a query vector.

A central design parameter of any such system is the dimensionality of the vector space. Dimensionality determines how much information each representation can encode, but it also governs the cost of storing and searching the representations. In practice, this parameter is frequently chosen by convention—values such as 100, 200, or 300 are inherited from influential prior systems—rather than selected through measurement on the task at hand.

This convention-driven approach is problematic because the costs of dimensionality are not free. Every additional dimension increases the memory required to store the embedding matrix, increases the time required to compute similarities during retrieval, and increases the number of parameters that must be estimated from finite training data. The benefits, by contrast, are subject to diminishing returns: beyond some point, additional dimensions encode little additional useful structure and may instead fit noise in the training corpus.

This paper investigates the trade-off empirically. The central question is direct: as embedding dimensionality increases, how do retrieval quality, query latency, and memory footprint change, and where does the most favourable balance lie? To answer it, embeddings were trained at a range of dimensions on a fixed corpus and evaluated on independent, standard benchmarks, with computational cost measured directly rather than assumed.

The contribution of this work is a clean, reproducible characterization of the dimensionality trade-off on a modest corpus, together with the observation that quality saturates well before cost does, and that one of the two quality benchmarks studied actually declines at higher dimensions.

II. BACKGROUND

A. Distributional Representations

The representations studied here rest on the distributional hypothesis, the long-standing observation that words occurring in similar contexts tend to have similar meanings. Methods that learn from co-occurrence statistics turn this hypothesis into vectors: each word is assigned a position such that words sharing many contexts are placed near one another. The skip-gram model used in this study learns these positions by training a shallow network to predict the context words that surround a given target word, with the learned hidden-layer weights serving as the embeddings.

B. The Role of Dimensionality

The number of dimensions in the embedding space is the capacity of the representation. Too few dimensions force unrelated concepts to share directions, blurring distinctions the model should preserve. Too many dimensions provide capacity that the available training data cannot reliably fill, so that some directions come to encode corpus-specific noise rather than transferable meaning. Between these extremes lies a region where capacity matches the information actually present in the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

data. Locating that region for a given corpus is the practical problem this paper addresses.

C. Semantic Retrieval

Once embeddings are available, semantic retrieval proceeds by embedding a query and ranking candidate items by their cosine similarity to it. Fig. 1 illustrates this pipeline. The dimensionality of the vectors enters this pipeline at every stage: it sets the size of each stored vector, the cost of each similarity computation, and the quality of the ranking that results.

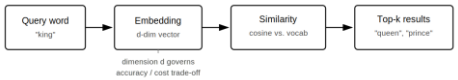

Fig. 1.
The semantic retrieval pipeline. The embedding dimension d determines both the quality of the ranking and the storage and computation cost at each stage.

III. METHODOLOGY

A. Corpus

Embeddings were trained on the Gutenberg corpus distributed with the Natural Language Toolkit, a collection of eighteen public-domain literary works including novels by Jane Austen, Herman Melville's Moby Dick, several plays by Shakespeare, the King James Bible, and assorted poetry and children's literature. The raw text was lower-cased, segmented into pseudo-sentences at sentence-terminating punctuation, and tokenized into alphabetic tokens. After filtering, the training corpus comprised 137,880 sentences and 2,123,697 tokens. A minimum count threshold of ten occurrences yielded a vocabulary of 9,979 word types, which was held identical across all dimensions so that vocabulary size could not confound the comparison.

B. Embedding Training

Skip-gram Word2Vec models were trained at five dimensions: 25, 50, 100, 200, and 300. All other hyperparameters were held constant—a context window of five words, five training epochs, and a fixed random seed—so that dimensionality was the only variable that changed between conditions. Holding the seed and all incidental settings fixed ensures that observed differences are attributable to dimensionality rather than to training variance.

C. Evaluation

Two independent and widely used intrinsic benchmarks were employed so that no single metric would dominate the conclusion. The first is the Mikolov word-analogy task, which poses questions of the form “a is to b as c is to ?” and is scored by the fraction of analogies for which the correct answer is the nearest neighbour of the predicted vector. The second is the WordSim-353 task, in which human annotators rated the relatedness of word pairs; performance is the Spearman rank correlation between human ratings and embedding cosine similarities. Because the two benchmarks probe different aspects of representation quality—relational structure versus graded similarity—agreement between them lends confidence to any conclusion they share.

D. Cost Measurement

Query latency was measured as the mean wall-clock time to retrieve the ten nearest neighbours of a query word, averaged over two hundred randomly selected frequent query words after a warm-up query. Memory footprint was computed as the exact size in bytes of the embedding matrix. Both quantities were measured on the same machine under identical conditions for every dimension.

IV. RESULTS

Table I reports the complete results across all five dimensions. Each row corresponds to one embedding model and lists both quality benchmarks alongside the measured cost quantities.

Dim.	Analogy (%)	WS-353 ρ	Latency (ms)	Mem. (MB)
25	6.06	0.241	0.101	0.95
50	8.38	0.268	0.155	1.90
100	8.79	0.233	0.229	3.81
200	8.45	0.212	0.371	7.61
300	8.84	0.212	0.567	11.42

TABLE I. *Quality and cost across embedding dimensions.*

A. Quality Saturates Early

Fig. 2 plots the two quality benchmarks against dimension. The analogy accuracy rises steeply from 6.06% at dimension 25 to 8.79% at dimension 100—capturing the great majority of the total improvement—and then flattens, with dimension 200 actually dipping slightly to 8.45% before dimension 300 reaches 8.84%. In other words, tripling the dimensionality from 100 to 300 changed analogy accuracy by only five hundredths of a percentage point. The similarity benchmark tells an even sharper story: WordSim-353 correlation peaks at dimension 50 ($\rho = 0.268$) and then declines monotonically to 0.212 at dimensions 200 and 300. The two benchmarks therefore agree that quality saturates at low-to-moderate dimensionality, and the similarity benchmark suggests that excess dimensionality is mildly harmful, consistent with higher-dimensional models fitting corpus-specific noise.

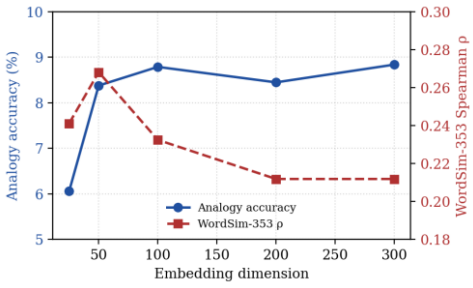

Fig. 2.
Retrieval quality versus embedding dimension on two independent benchmarks. Both saturate early; the similarity benchmark peaks at dimension 50 and declines thereafter.

B. Cost Grows Linearly

Fig. 3 plots the two cost quantities against dimension. In contrast to the saturating quality curves, both latency and memory grow strictly linearly with dimension, as expected from the structure of the computation: storing and comparing vectors is proportional to their length. Mean query latency rose from 0.101 ms at dimension 25 to 0.567 ms at dimension 300,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

a factor of 5.6, while the embedding matrix grew from 0.95 MB to 11.42 MB, a factor of 12. These costs were paid in full regardless of whether they purchased any improvement in quality.

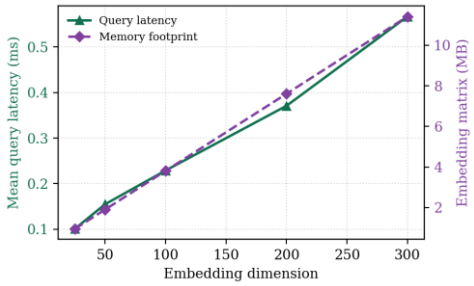

Fig. 3.
Query latency and memory footprint versus embedding dimension. Both grow linearly, in contrast to the saturating quality curves of Fig. 2.

C. The Efficiency Frontier

The practical consequence of one curve saturating while the other grows linearly is that accuracy-per-resource falls steadily as dimensionality increases. Fig. 4 makes this explicit by plotting analogy accuracy per megabyte of memory. The smallest model is by far the most efficient on this measure, and efficiency declines at every step toward higher dimensionality. Where retrieval quality is the goal and resources are finite, the data favour dimensions in the range of 50 to 100: dimension 50 captures most of the available quality at a fifth of the memory of dimension 300, while dimension 100 reaches essentially the maximum analogy accuracy at a third of that memory.

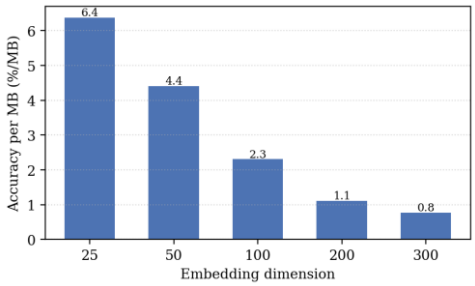

Fig. 4.
Analogy accuracy per megabyte of memory. Efficiency is highest at low dimensionality and declines monotonically.

V. DISCUSSION

The findings carry a simple practical message: dimensionality should be measured against the task and corpus rather than inherited from convention. On the corpus studied, conventional choices such as 200 or 300 dimensions delivered no quality advantage over 100 while imposing two-to-three times the cost, and on the similarity benchmark they performed worse than a model a sixth their size.

The result that one quality benchmark declines at higher dimensions deserves emphasis. It is consistent with the view that representation capacity must be matched to the information available in the data: when a corpus is modest, high-dimensional models have more capacity than the data can constrain, and the surplus directions absorb idiosyncratic co-occurrence patterns that do not generalize. This suggests that the optimal dimensionality is not a universal constant but scales with the richness of the training corpus, and that practitioners working with limited data should be especially wary of large embedding sizes.

It is important to delimit the scope of these claims. The study uses a single modest corpus, classical word-level embeddings rather than contextual representations, and intrinsic benchmarks rather than a deployed end-to-end retrieval task. The specific saturation point—near dimension 100 here—should therefore not be read as a universal recommendation; the transferable findings are the qualitative shape of the trade-off and the methodology for locating the saturation point on any given corpus.

VI. CONCLUSION AND FUTURE WORK

This paper presented a controlled empirical study of how word-embedding dimensionality affects retrieval quality, query latency, and memory footprint. Across two independent benchmarks, quality was found to saturate at low-to-moderate dimensionality—by dimension 100 for analogies and as early as dimension 50 for similarity—while latency and memory grew linearly without bound. The highest-dimensional model studied therefore cost an order of magnitude more in memory than the smallest for no measurable quality gain, and underperformed a much smaller model on the similarity task. The most favourable accuracy-per-resource trade-off lay at low dimensionality.

Several extensions follow naturally. The dependence of the saturation point on corpus size could be characterized directly by repeating the experiment across corpora of increasing size, testing the conjecture that optimal dimensionality scales with data richness. The study could be broadened to contextual sentence embeddings and to a deployed end-to-end retrieval task with relevance judgements, to confirm that the intrinsic findings translate to downstream performance. Finally, post-hoc dimensionality-reduction techniques could be compared against training at the target dimension, to determine whether the saturation behaviour is a property of the training process or of the representation itself.

REFERENCES

[1]T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. Learning Representations (ICLR) Workshop, 2013.

[2]T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 3111–3119.

[3]J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[4]Z. S. Harris, “Distributional structure,” Word, vol. 10, no. 2–3, pp. 146–162, 1954.

[5]L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, “Placing search in context: The concept revisited,” ACM Trans. Information Systems, vol. 20, no. 1, pp. 116–131, 2002.

[6]R. Řehůřek and P. Sojka, “Software framework for topic modelling with large corpora,” in Proc. LREC Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.

[7]S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA: O'Reilly Media, 2009.