Visualizing High-Dimensional BERT Embeddings Using Radar Charts

Yannis Bendi-Ouis
Mnemosyne, Inria
Bordeaux University

Xavier Hinaut
Mnemosyne, Inria
Bordeaux University

Abstract

This paper presents a novel approach to visualize high-dimensional BERT embeddings using radar charts. By arranging the 768 dimensions of BERT embeddings to maximize their absolute correlation, and applying gaussian filters within the arranged dimensions, we provide a visualization without reducing dimensionality. Despite the difficulty of understanding the meaning of each representation, this new method offers new perspectives on the visualization of the semantic relationships within the embeddings.

1 Introduction

Word embeddings [6], specifically those derived from BERT models [2], offer a rich, high-dimensional space capturing nuanced semantic relationships. Traditional visualization techniques like PCA [4] and t-SNE [5] simplify these embeddings to lower dimensions, often losing significant information. Our approach, in contrast, visualizes these embeddings in their original dimensionality using radar charts, aiming to arrange dimensions based on their correlation (Fig 1), which preserves more semantic subtlety.

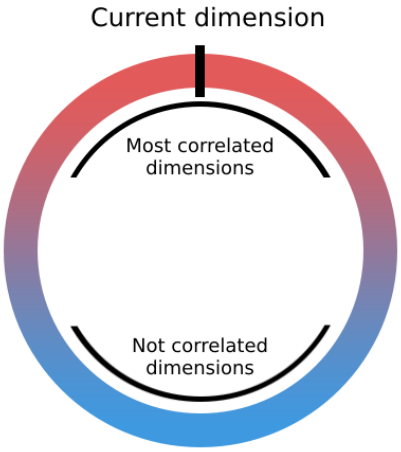

Figure 1: Arrangement of dimensions around the circle.

Preprint. Under review.

2 Methodology

2.1 Problem Complexity and Approach

In our study, we tackle the challenge of effectively displaying 768-dimensional BERT embeddings [2] on a radar chart. The goal is to find a circular path among these dimensions that maximizes the correlation between each dimension and its closest neighbors while minimizing the correlation between each dimension and its diametrically opposite dimensions on the chart.

This task is equivalent to the well-known NP-Complete problem, the Traveling Salesman Problem [7], where the aim is to determine the shortest possible path visiting a set of points. To solve this, we employ a genetic algorithm [8], which we will detail below, to navigate through the complex, high-dimensional space and optimize the arrangement of dimensions for the radar chart visualization.

2.2 Dimension Analysis and Treatment

In our analysis of the dimension correlations within BERT embeddings [2], we focused on the absolute values of these correlations, with most falling between 0.3 and 0.4. We focused on the absolute correlation because, in the context of embeddings, both positive and negative correlations are equally significant indicators of a relationship between dimensions.

By considering the absolute values, we effectively capture the strength of these relationships without the ambiguity associated with the direction of correlation. This is akin to acknowledging that in the multidimensional space of embeddings, the concepts of 'positive' and 'negative' are not inherently meaningful - much like the absence of a universal 'up' or 'down' in space.

2.3 Genetic Algorithm Components

2.3.1 Fitness Calculation Rule

In our genetic algorithm, the fitness function evaluates the suitability of a path in the embedding space. To compute the fitness, we sum the distances between each dimension and every other dimension, resulting in an N^2 complexity where N is the number of dimensions.

This distance, based on the formula $f(c_{i,j}) = \pi - \frac{\pi c_{i,j}}{C_{max}}$ (Fig. 2), where $c_{i,j}$ is the correlation between dimension i and j, is what we aim to minimize.

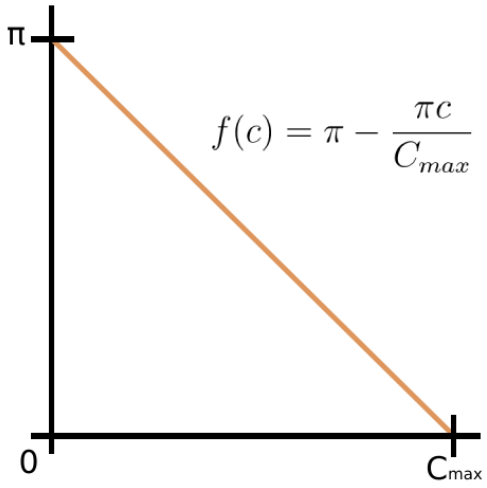

Figure 2: Desired distance based on correlation.

The function essentially compute the desired distance between two dimensions so that those with high correlation (close to C_{max}) are closed on the circle (distance close to 0). And those with low correlation (close to 0) are positioned on the opposite of the circle (distance close to pi).

In doing so, it aims to minimize the overall distance error, effectively arranging the dimensions in a manner that optimizes their correlation relationships.

2.3.2 Mutation Mechanism

The mutation process in our genetic algorithm involves strategically relocating certain elements within a path. This is done by first selecting elements at random based on a mutation rate and then reinserting them into positions where they are more closely aligned with adjacent dimensions in terms of correlation.

2.3.3 Crossover Mechanism

The crossover function combines segments from two parent paths to generate a new path. It selects a random subpath from a parent, then fills the remaining positions with the path from the second parent, without duplicating a dimension, and preserving the order found in the second parent.

2.4 Dimension Reversibility

In our approach, a dimension's sign is reversed if it is negatively correlated with its preceding dimension in the path. This ensures that dimensions with negative correlations move in tandem, increasing and decreasing together.

2.5 Gaussian Filter Application

We apply a Gaussian filter [9] to each point in the embedding to smooth out the variations across dimensions. The filter considers a specified number of dimensions before and after the current one, encompassing a total of $2N + 1$ dimensions. This smoothing process helps in reducing noise and making the overall patterns in the data more discernible (Fig 3).

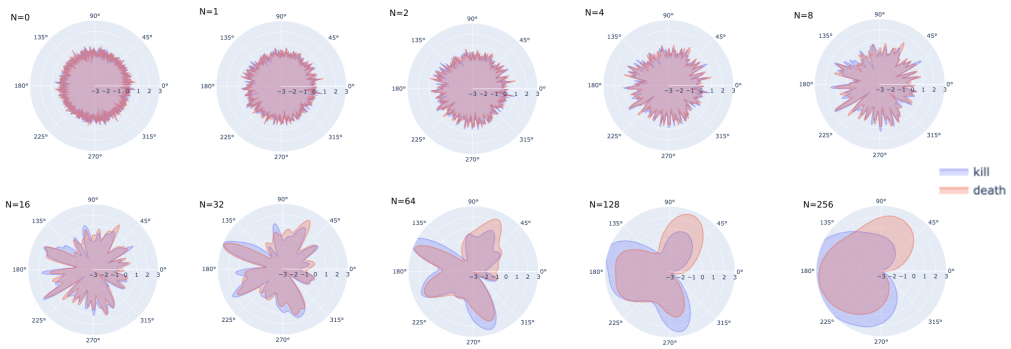

Figure 3: Evolution of the shape in function of the Gaussian size.

2.6 Sampling Methods and Correlation Analysis in BERT Embeddings

To calculate the correlations between BERT embedding's [2] dimensions, we first needed to gather samples. We experimented with three different data collection methods:

1. Using all 30.000+ tokens known by BERT and its tokenizer.
2. Extracting tokens from a set of diverse texts.
3. Extracting tokens from a single long text: "The Little Prince" by Antoine de Saint-Exupéry.

Contrary to our hypothesis, all methods yielded similar results, with the majority of correlations falling between 0.3 and 0.4.

3 Experiments and Results

3.1 Comparison of the Shape Obtained for Common Words

After around 200 to 350 generations, the genetic algorithm stops finding a better path, suggesting an optimal arrangement of dimensions. This arrangement was used to generate the radar chart from BERT Embeddings (Fig. 4).

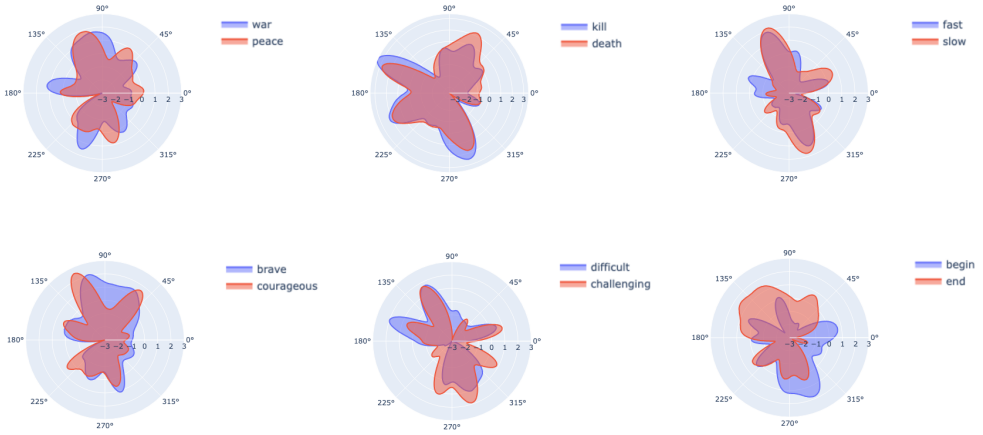

Figure 4: Superposition of common word's shapes obtained by our methods.

As can be seen in Fig. 4, some words with close semantic relationships "overlap", while others with opposite semantic meanings seem to retain an overlap with more major differences.

That said, it's not always easy to interpret or predict these results. Indeed, some words that are expected to be radically different may have shapes that seem to overlap on several aspects, and conversely, words that may seem very similar to us may have distinct shapes.

3.2 Contextual Influence on Embedding Evolution

In our study, we conducted two distinct experiments to see the evolution of the embedding's shape in varying contexts. The first experiment focused on the evolution of the representation of a single word as we incrementally added context before and after it (Fig. 5). The second experiment delved into the evolution of sentence embeddings, observing how the representation changes with the sequential addition of words to the end of a sentence (Fig. 6).

These explorations revealed significant shifts in the embeddings with varying context, underscoring the sensitivity of word representations to surrounding text, and the impact that some words can have on the embedding of a whole sentence. However, despite these noticeable changes, extracting clear semantic interpretations from these evolving representations posed a substantial challenge.

We also examined the effect of adding one word to the end of a sentence (Fig. 7). Our expectation was that the representation of a sentence at time t would resemble that at time $t-1$, but with modifications leaning towards the word at time t. Interestingly, we observed the opposite: the shape tends to contract in areas where significant peaks are observed in the word at time t, and expand in areas with troughs.

3.3 Cosine Similarity Analysis

In an effort to explore the effectiveness of our visualization method in capturing semantic relationships, we conducted an analysis of cosine similarity [10] between word pairs. To this end, ChatGPT generated two lists of 40 word pairs: one containing semantically close words and the other containing semantically distant words. A sample of these pairs is presented below in Table 1.

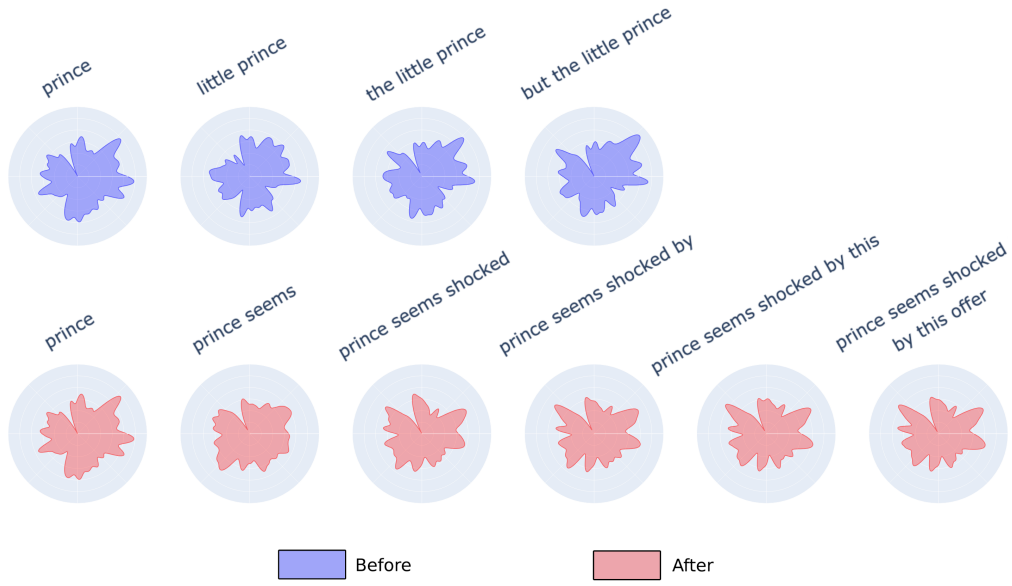

Figure 5: Evolution of the shape of the word "prince" in function of its context (before and after).

Word 1	Word 2	Before	After
joy	happiness	39.0	36.0
angry	furious	55.0	70.0
laugh	giggle	49.0	68.0
jump	leap	34.0	14.0
run	sprint	47.0	58.0
fear	terror	42.0	45.0
sad	unhappy	34.0	27.0
build	construct	49.0	59.0
breeze	wind	49.0	59.0
chilly	cold	36.0	41.0
poetry	insect	13.0	4.0
happiness	garbage	34.0	35.0
science	feather	34.0	44.0
philosophy	basketball	38.0	71.0
thunder	sorrow	10.0	9.0
novel	cactus	31.0	1.0
ocean	lantern	56.0	47.0
painting	hunger	24.0	19.0
dream	ladder	25.0	4.0
galaxy	ant	27.0	52.0

Table 1: Cosine similarity of semantically close and distant word pairs before and after Gaussian smoothing.

Upon analyzing these pairs, we focused on the cosine similarity values before and after applying Gaussian smoothing [9] to the embeddings. The results, derived from a total of 80 combinations, are visually represented in Fig. 8. From these observations, the average cosine similarities were calculated, yielding the results in Table 2. These results indicate a slight increase in average similarity for both close and distant pairs after the application of Gaussian smoothing.

Figure 6: Evolution of the shape for a whole sentence, word by word.

Pair Type	Before Smoothing	After Smoothing
Close Pairs	0.440	0.453
Distant Pairs	0.365	0.373

Table 2: Average cosine similarities for close and distant word pairs before and after Gaussian smoothing.

4 Conclusion

This novel visualization method offers a nuanced approach to understanding BERT embeddings [2]. It provides a representation that is more faithful to the true multi-dimensional nature of these embeddings than traditional dimensionality reduction techniques like PCA [4] or t-SNE [5], enabling a clearer perception of the proximity between words or phrases as interpreted by BERT.

However, this method still falls short in unraveling the hidden semantics within BERT's complex embeddings. It does not necessarily offer a clearer understanding of the reasons behind the closeness or distance between words or phrases according to BERT's metrics.

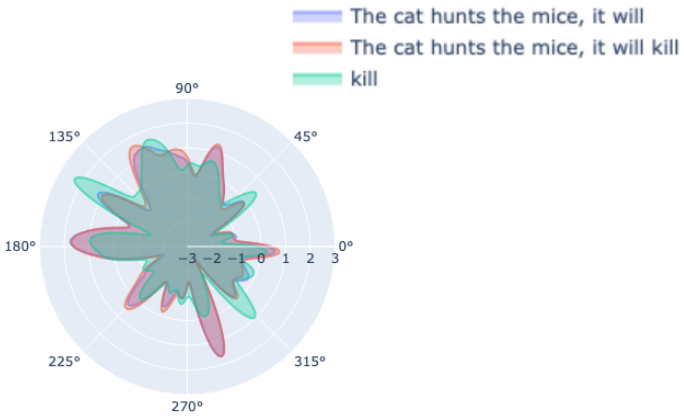

Figure 7: A sentence compared to itself plus next word compared to next word.

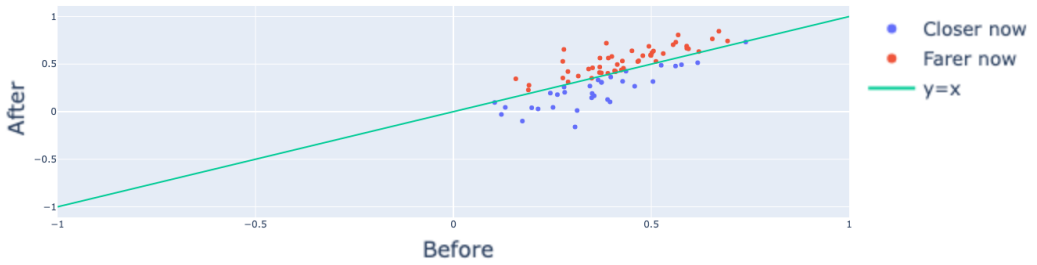

Figure 8: Cosine similarity before/after comparison.

Despite these limitations, our method retains certain properties of BERT embeddings, particularly in terms of cosine similarity [10] between vectors, which seems to be partially preserved, though this is slightly altered by the Gaussian filtering [9].

Interestingly, the overlay of shapes in our visualizations does not always align with intuitive expectations. Words that seem semantically close may appear distant in this representation and vice versa, suggesting a disparity between human intuition and BERT's embedding logic.

Yet, this method opens up new avenues for interpreting the dimensions resulting from BERT and the potential modeling thereof. Additionally, the artistic aspect of this new representation may pave the way for art-science intersections in the realm of Large Language Models (LLMs).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need". arXiv:1706.03762v7, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2, 2018.

[3] J. Li, X. Chen, E. Hovy, et al., "Visualizing and Understanding Neural Models in NLP". arXiv:1506.01066v2, 2016.

[4] I. T. Jolliffe, "Principal Component Analysis". Springer Series in Statistics, 2002.

[5] L. van der Maaten, G. Hinton, "Visualizing Data using t-SNE". Journal of Machine Learning Research, 2579–2605, 2008.

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781v3, 2013.

[7] D. Applegate, R. Bixby, V. Chvátal, W. Cook, "The Traveling Salesman Problem: A Computational Study". Princeton University Press, 2006.

[8] D. E. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning". Addison-Wesley, 1989.

[9] T. Lindeberg, "Scale-space for discrete signals". IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(3), 234-254, 1990.

[10] A. Singhal, "Modern Information Retrieval: A Brief Overview". IEEE Data Engineering Bulletin, 24(4), 35-43, 2001.