On Geodesic Distances and Contextual Embedding Compression for Text Classification

Rishi Jha* and **Kai Mihata***
Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
{rjha01, kaim2}@cs.washington.edu

Abstract

In some memory-constrained settings like IoT devices and over-the-network data pipelines, it can be advantageous to have smaller contextual embeddings. We investigate the efficacy of projecting contextual embedding data (BERT) onto a manifold, and using nonlinear dimensionality reduction techniques to compress these embeddings. In particular, we propose a novel post-processing approach, applying a combination of Isomap and PCA. We find that the geodesic distance estimations, estimates of the shortest path on a Riemannian manifold, from Isomap's k-Nearest Neighbors graph bolstered the performance of the compressed embeddings to be comparable to the original BERT embeddings. On one dataset, we find that despite a 12-fold dimensionality reduction, the compressed embeddings performed within 0.1% of the original BERT embeddings on a downstream classification task. In addition, we find that this approach works particularly well on tasks reliant on syntactic data, when compared with linear dimensionality reduction. These results show promise for a novel geometric approach to achieve lower dimensional text embeddings from existing transformers and pave the way for data-specific and application-specific embedding compressions.

1 Introduction

Contextual embeddings, like those BERT (Devlin et al., 2019) generates, improve on non-contextual word embeddings by providing contextual semantics to the real-valued representation of a text. Although these models have been shown to achieve state-of-the-art performance on most NLP tasks, they are notably expensive to train. To help combat this, as mentioned by May et al. (2019), model compression techniques like data quantization (Gong et al., 2014), model pruning (Han et al., 2016), and knowledge distillation (Sanh et al., 2019, Hinton et al., 2015) have been developed. However, at 768 dimensions, the embeddings themselves can be prohibitively large for some tasks and settings.

Smaller embeddings both enable more compact data sizes in storage-constrained settings and over-the-air data pipelines, and help lower the requisite memory for using the embeddings for downstream tasks. For non-contextual word embeddings, Ling et al. (2016) note that loading matrices can take multiple gigabytes of memory, a prohibitively large amount for some phones and IoT devices. While contextual embeddings are smaller, downstream models will face similar headwind for large corpora.

Although there has been more extensive study in the efficacy of compressing non-contextual word embeddings (Raunak et al., 2019, Mu and Viswanath, 2018), to the best of our knowledge few contextual embedding compression post-processing approaches have been proposed (Li and Eisner, 2019). In their work, Li and Eisner (2019) propose the Variational Information Bottleneck, an autoencoder to create smaller, task specific embeddings for different languages. While effective, the computational expense of additional training loops is not appropriate for some memory constrained applications.

Our approach more closely mirrors the work of Raunak et al. (2019) who propose a Principal Component Analysis (PCA)-based post-processing algorithm to lower the dimensionality of non-contextual word embeddings. They find that they can replicate, or, in some cases, increase the performance of the original embeddings. One limitation to this approach is the lack of support for nonlinear data patterns. Nonlinear dimensionality reductions, like the Isomap shown in Figure 1, can pick up on latent textual features that evade linear algorithms like PCA. To achieve this nonlinearity, we extend this approach to contextual embeddings, adding in

*Equal contribution

Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), pages 144–149

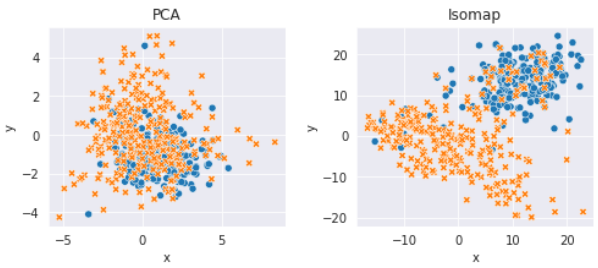

Figure 1: Visualization of two-dimensional PCA and Isomap compressions based on BERT embeddings for the SMS-SPAM dataset (Almeida et al., 2013). Spam is represented by a blue dot and ham by an orange x. We see that for this dataset in two dimensions, the Isomap compression appears more linearly separable than the PCA compression, making classification easier for the former.

additional geodesic distance information via the Isomap algorithm (Tenenbaum et al., 2000). To the best of our knowledge, the application of graph-based techniques to reduce the dimensionality of contextual embeddings is novel.

The goal of this paper is not to compete with state-of-the-art models, but, rather, (1) to show that 12-fold dimensionality reductions of contextual embeddings can, in some settings, conserve much of the original performance, (2) to illustrate the efficacy of geodesic similarity metrics in improving the downstream performance of contextual embedding compressions, and (3) propose the creation of more efficient, geodesic-distance-based transformer architectures. In particular, our main result is showing that a 64-dimensional concatenation of compressed PCA and Isomap embeddings are comparable to the original BERT embeddings and outperform our PCA baseline. We attribute this success to the locality data preserved by the k-Nearest Neighbors (k-NN) graph generated by the Isomap algorithm.

2 Related Work

As best we know, there is very little literature regarding the intersection of contextual embedding compression and geodesic distances. Most of the existing work in related spaces deals with non-contextual word embeddings. Despite the rapid growth in the popularity of transformers, these embeddings still retain popularity.

For non-contextual word embeddings, Mu and Viswanath (2018) propose a post-processing algorithm that projects embedded data away from the dominant principal components, in order to greater differentiate the data. Raunak et al. (2019) expand on this algorithm by combining it with PCA reductions. Both approaches are effective, but, are limited to linear dimensionality reductions.

Some nonlinear approaches include Andrews (2016) and Li and Eisner (2019) who both use autoencoder-based compressions. Notably, the former only addresses non-contextual embeddings.

Meanwhile the usage of graphs in NLP is well established, but their usage in the compression of contextual embeddings is not well documented. Wiedemann et al. (2019) use a k-NN classification to achieve state-of-the-art word sense disambiguation. Their work makes clear the effectiveness of the k-NN approach in finding distinctions in hyper-localized data.

3 Method

With the goal of reducing the contextual embedding dimensionality, we first processed our data using a pre-trained, uncased BERT Base model. Then, we compressed the data to a lower dimension using both PCA and Isomap as described in Section 3.2. This method aims to capture as much information as possible from the original BERT embeddings while preserving graphical locality information and nonlinearities in the final contextual embeddings.

3.1 Isomap and Geodesic Distances

For this paper, to blend geodesic distance information and dimensionality reduction, we use Tenenbaum's Isomap (Tenenbaum et al., 2000). Isomap relies on a weighted neighborhood graph that allows for the inclusion of complex, nonlinear patterns in the data, unlike a linear algorithm like PCA. In specific, this graph is constructed so that the edges between each vertex (datapoint) and its k-nearest neighbors have weight corresponding to the pairwise Euclidean distance on a Riemannian manifold. Dijkstra's shortest path between two points then estimates their true geodesic distance.

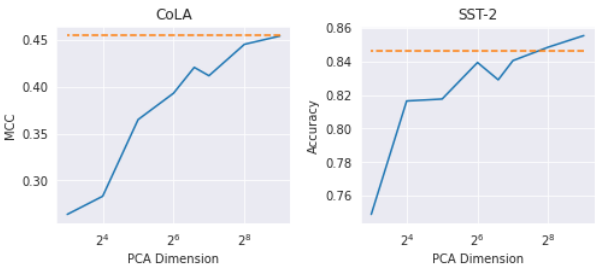

Figure 2: PCA baseline performance on CoLA (Warstadt et al., 2019) and SST-2 (Socher et al., 2013). PCA embedding performance by dimension is represented by the solid blue line. Regression at 768 dimensions is represented by an orange dashed line. On these two datasets, even at much smaller dimensionality, we see that PCA has comparable performance.

These geodesics are particularly useful for delineating points that are close in Euclidean space but not on a manifold i.e. similar BERT embeddings with different meanings.

If we assume the data follows a manifold, Isomap can exploit the Riemannian locality of these complex contextual embeddings. As Figure 1 shows, in some cases this is a good assumption to make since we are then able to dissect complex embeddings into near-linearly separable clusters. Notably, there are some limitations to this approach. If the manifold is sparse, i.e. there are few data points on certain regions of the manifold, or k is too small, the shortest path estimation of the geodesic distance can be unrepresentative of the true distance. On the contrary, if k is too large, Isomap overgeneralizes and loses its fine-grained estimation of the Riemannian surface.

Nonetheless, we hypothesize that these global geodesic distance approximations explain the empirical advantage Isomap has in our setting over other popular nonlinear dimensionality reduction techniques. Many alternatives, like Locally Linear Embeddings (Roweis and Saul, 2000) focus, instead, on preserving intra-neighborhood distance information that may not encompass inter-neighborhood relationships as Isomap does.

3.2 Our Approach

We applied our post-processing method to the BERT embeddings through three different dimensionality reductions. We used (1) PCA, (2) Isomap, and (3) a concatenation of embeddings from the two before training a small regression model on the embeddings. This approach aims to use linear and nonlinear dimensionality reduction techniques to best capture the data's geodesic locality information.

PCA. To compute linearly-reduced dimensionality embeddings, we used PCA to reduce the 768-dimensional BERT embeddings down to a number of components ranging from 16 to 256. While there are other linear dimensionality reduction techniques, PCA is a standard benchmark and empirically performed the best. These serve as a linear baseline for reduced dimension embeddings.

Isomap. To compute geodesic locality information, we post-processed our BERT embeddings with Isomap. The final Isomap embeddings ranged from 16 to 96 dimensions, all computed with 96 neighbors and Euclidean distance.

Concatenated Embeddings. To include features from both of these reductions, we combined an Isomap embedding with a PCA embedding to form concatenations of several dimensions. We experimented with ratios of PCA embedding size to Isomap embedding size from 0 to $\frac{1}{2}$ at $\frac{1}{8}$ intervals. We found that this ratio was the main determinant of relative accuracy, so for analysis we fixed the total dimension to 64.

4 Experiments and Results

We assess the results of these compression techniques on two text classification datasets. We provide the code for our experiments[1].

4.1 Data

We evaluate our method on two text classification tasks: the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019), a 10657 sentence binary classification dataset on grammatical correctness and the Stanford Sentiment Treebank v2 (SST-2) (Socher et al., 2013), a 70042 sentence

[1] https://github.com/kaimihata/geo-bert

Embedding	Isomap Dim	PCA Dim	CoLA	SST-2
PCA	N/A	64	0.339	0.842
Concatenation*	16	48	0.421	0.846
Concatenation	32	32	0.384	0.822
Concatenation	48	16	0.357	0.817
Isomap	64	N/A	0.332	0.814
BERT	N/A	N/A	0.455	0.847

Table 1: 64-dimensional embedding performance on CoLA (Warstadt et al., 2019) and SST-2 (Socher et al., 2013). CoLA is measured by Matthews correlation and SST-2 by accuracy. While Isomap did not perform the best outright, on these datasets we found that some inclusion of locality data proved meaningful. This shows the trade-off between locality information and performance mentioned in Section 4.4. The best 12-fold compression performance is asterisked.

binary (positive / negative) sentiment classification dataset.

For CoLA, we used the predefined, 9594 datapoint train set and for SST-2, we used the first 8000 samples of their training set to construct ours due to computational limitations. For testing and evaluation, we used the corresponding datasets defined by GLUE (Wang et al., 2018). In addition, for all of our evaluations, we used the same pre-trained BERT embeddings for consistency.

4.2 Training and Evaluation

All of these post-processed embeddings, as well as the BERT embeddings, were trained on a downstream regression model consisting of one hidden layer (64 dim) with ReLU activation, a learning rate of 1×10^{-4}, and were optimized via ADAM (Kingma and Ba, 2015). The BERT embeddings are used as a baseline for comparison.

To evaluate our embeddings on CoLA and SST-2, we used their GLUE-defined metrics of Matthews correlation and validation accuracy, respectively. For each embedding experiment, our procedure consisted of running our post-processing method on the BERT embeddings then training the downstream model. Each reported metric is the average of three of these procedures.

4.3 Baseline Comparison

Agnostic of post-processing algorithm, we found reduced-dimensionality embeddings were competitive with the original embeddings. Although smaller reduction factors, understandably, performed better, we found that even when reduced by a factor as large as 12, our PCA embeddings experienced small losses in performance on both datasets (Figure 2). To demonstrate the effect of the inclusion of locality data, we picked an embedding size of 64 dimensions (a reduction factor of 12) to balance embedding size and performance for our main experiment.

In comparison to our 768-dimensional baseline, at 64 dimensions, the best reduction results were within 7.5% and 0.1% for CoLA and SST-2, respectively (Table 1). These results show that with or without the presence of locality data, compressed embeddings can perform comparably to the original embeddings.

4.4 Locality Information Trade-off

As shown in Table 1, on neither dataset did the fully PCA or Isomap embeddings perform the best. The best performer was, instead, a combination of these two approaches. This indicates that there must exist a trade-off on the effectiveness of locality data. While without locality data, the embedding obviously misses out on geodesic relationships, too much locality information may replace more useful features that the PCA embeddings extract. Just as the quality of the geodesic distance estimations rely on how well the data fits the underlying manifold, as discussed in Section 3, so, too, does its effectiveness. To explain this phenomenon, we hypothesize that the addition of small amounts of locality data bolsters performance by describing the geodesic relationships without drowning out important syntactic and semantic information provided by PCA.

4.5 Task-Specific Locality

While the best reduction consisted of a concatenation of 16-dimensional Isomap and 48-dimensional PCA embeddings, whether the other concatenations performed better than our PCA baseline was dependent on the task. For CoLA, we found that all three concatenated embeddings performed better than PCA, whereas for SST-2, only the top perform-

ing concatenated embedding beat out our baseline. To describe this disparity we look towards the nature of the datasets and tasks. Notably, CoLA requires models to identify proper grammar, a syntactic task, while SST-2 requires models to understand the sentiment of sentences, a semantic task. Syntactic data often has some intrinsic structure to it, and perhaps our manifold approach encompasses this information well. Based on this result, exploring this distinction could be an exciting avenue for further study.

5 Conclusions and Future Work

We present a novel approach for compressing BERT embeddings into effective lower dimension representations. Our method shows promise for the inclusion of geodesic locality information in transformers and future compression methods. We hope our results lead to more work investigating the geometric structure of transformer embeddings and developing more computationally efficient NLP training pipelines. To further this work, we plan to investigate the efficacy of (1) other graph dimensionality reduction techniques, (2) non-Euclidean distance metrics, and (3) our approach on different transformers. In addition, we would like to investigate whether datasets for other tasks can be effectively projected onto a manifold.

Acknowledgments

We would like to thank Raunak Kumar, Rohan Jha, and our three reviewers for their thoughtful and thorough comments on improving our paper. In addition, we would like to thank our deep learning professor Joseph Redmon for inspiring this project.

References

Tiago Almeida, Jose Gomez Hidalgo, and Tiago Silva. 2013. Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*, 2:1–18.

Martin Andrews. 2016. Compressing word embeddings. In *Neural Information Processing*, pages 413–422, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing Deep Convolutional Networks using Vector Quantization. *arXiv e-prints*, page arXiv:1412.6115.

Song Han, Huizi Mao, and William J Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *International Conference on Learning Representations (ICLR)*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2744–2754, Hong Kong, China. Association for Computational Linguistics.

Shaoshi Ling, Yangqiu Song, and Dan Roth. 2016. Word embeddings with limited memory. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 387–392, Berlin, Germany. Association for Computational Linguistics.

Avner May, Jian Zhang, Tri Dao, and Christopher Ré. 2019. On the Downstream Performance of Compressed Word Embeddings. *arXiv e-prints*, page arXiv:1909.01264.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. Association for Computational Linguistics.

Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.