Lossy Text Compressor

Swati Aggrawal

Link to the Github Repository

Abstract

This study explores lossy text compression using deep learning-based models, focusing on the impact of quantization on text reconstruction. Semantic embeddings of text are created using the GTR-T5 Base model and compressed using several quantization levels (2, 4, 6, 8, 10, 12, 14, 16 bits). The text is reconstructed from compressed embeddings using the Vec2Text model. The quality of reconstruction is evaluated using metrics like ROUGE-L, BLEU, and Cosine Similarity. The findings show that 6–8 bit quantization offers the optimal trade-off between structural integrity and meaning retention, with higher levels leading to significant losses.

1. Introduction

A key component of data transmission and storage is text compression, which aims to minimise the size of text representations without sacrificing meaning. Conventional techniques include lossy compression (e.g., vector quantization, embeddings), which minimises redundancy at the expense of some information loss, and lossless compression (e.g., Huffman coding (Kadhim et al., 2024), Lempel-Ziv (Ziv & Lempel, 1977)), which precisely reconstructs the original text. Research on the use of lossy techniques in text is still ongoing, despite the fact that they are often employed in speech and image compression.

This project investigates deep learning-based text reconstruction and text embeddings for lossy text compression. We apply the GTR-T5 Base (Ni et al., 2022) model to build dense vector representations of input text, capturing its semantic core. These embeddings undergo compression by quantization, decreasing precision to lower bit levels, hence drastically reducing storage requirements. We use Vec2Text (Morris et al., 2023), a transformer-based model trained to produce readable text from embeddings, to reconstruct the original text.

We employ three metrics for assessment to determine how compression affects text reconstruction:

- Cosine Similarity is used to determine how closely the original and reconstructed embeddings match.
- Based on n-gram precision, the BLEU Score is used to assess the accuracy of the reconstructed text.
- To evaluate the sequence-level overlap between the generated and original text, compute the ROUGE-L Score.

This study shows that modern embedding and generation models impact lossy text compression by analysing the impact of compression on reconstructed text using a variety of metrics.

2. Related Work

To enhance data processing, transmission, and storage efficiency, compression techniques are crucial. These techniques are broadly categorized into lossless and lossy compression. Lossless methods, such as Huffman coding (Kadhim et al., 2024), Run-Length Encoding (RLE) (Fiergolla & Wolf, 2021), and Lempel-Ziv-Welch (LZW) (Ziv & Lempel, 1977), ensure exact data recovery, making them suitable for accuracy-dependent applications like database management and file storage.

On the other hand, lossy compression achieves higher compression ratios by discarding some data, commonly used in multimedia formats like MP3 and JPEG. Witten et al. (1994) emphasize the trade-off between compression efficiency and text quality in lossy text compression, which has been underexplored. Palaniappan & Latifi (2007) discuss character-based techniques such as Dropped Vowels (DOV) and Letter Mapping (LMP) to improve compression for non-critical applications.

Lossy methods like quantization, pruning, and dimensionality reduction reduce memory usage but can impact data integrity. Quantization, especially, converts continuous

Email: Swati Aggrawal <aggrawal.2118031@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

data to discrete levels, lowering precision while maintaining essential data representation.

3. Methodology

As outlined in the following sections, the study's approach consists of generating text embeddings, compressing the text using quantization, reconstructing the text from compressed embeddings, and assessing the reconstruction quality using a variety of metrics.

3.1. Textual Embeddings Using Pre-trained Models

The pre-trained GTR-T5 model (Ni et al., 2022), a T5 variation intended for text repair and reconstruction through a transformer architecture, has been used. It is effective, integrates with Sentence-Transformers, and is pre-trained. It can produce rich and accurate semantic representations of textual data. Tokenizing the input text into smaller subword units and then encoding it together allows the model to generate an embedding for every sentence. This is accomplished by mean pooling across the last layer of the model's hidden states, guaranteeing a thorough text representation.

3.2. Quantization for Compression

While maintaining crucial information, quantization lowers embedding precision for efficient storage. By normalizing, scaling, and converting embeddings to lower precision (uint8), it recovers approximate values, and dequantization returns them to their original range. Among the steps are:

- **Normalization:** [0, 1] is the range of values for embeddings.
- **Quantization:** Reduces precision by rounding and scaling embeddings.
- **Dequantization:** The original range of quantized data is rescaled.

3.3. Text Reconstruction

Following quantization and dequantization, the embeddings are fed into the Vec2Text model (Morris et al., 2023), which uses the compressed embedding representation to recreate the original text. This makes it possible to examine the effects of compression on text recoverability.

3.4. Reconstruction Quality Assessment Metrics

The following important metrics are used to evaluate the quality of the reconstructed text:

1. **Cosine Similarity :** Measures the alignment between the original and reconstructed embeddings.

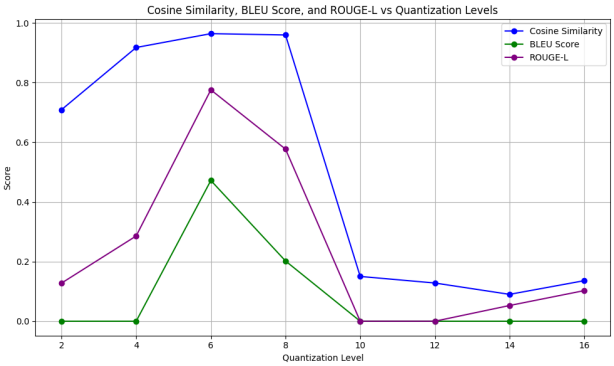

Figure 1. This graph shows the values of Cosine Similarity, BLEU, and ROUGE scores for each quantization level, highlighting their trends.

2. **BLEU Score :** Evaluates the overlap of n-grams between the original and reconstructed text.
3. **ROUGE-L Score :** Measures the longest common subsequence (LCS) overlap to capture structural similarity between the original and reconstructed text.

4. Results

Text reconstruction quality is affected by different quantization levels (2, 4, 6, 8, 10, 12, 14, 16 bits), as demonstrated by the project's results. Cosine similarity remains high at 6 (0.963949) and 8 (0.959555) bits, indicating strong semantic preservation, but decreases significantly at higher quantization levels, showing a loss in meaning. The BLEU score is very low at higher levels (10-16 bits), suggesting minimal overlap with the original n-grams and a loss of structural fluency. The ROUGE-L score is highest at 6 bits (0.775510), indicating good structural similarity, but decreases at higher levels, reflecting the loss of structural details due to lossy compression. Figure 1 illustrates the results.

5. Discussion and Conclusions

The results show that 6-bit and 8-bit quantization levels achieve the highest accuracy, as indicated by cosine similarity, preserving the core meaning of the text. Beyond these levels, reconstruction quality declines rapidly. Due to lossy compression, BLEU and ROUGE-L scores remain low, reflecting structural loss. However, the high cosine similarity at 6 and 8 bits confirms that essential semantic information is retained. Thus, these levels offer the best trade-off between meaning preservation and structural integrity. Future research could explore hybrid quantization, task-specific fine-tuning, and analyzing quantization effects across various text types and applications to enhance reconstruction quality.

References

Fiergolla, S. and Wolf, P. Improving run length encoding by preprocessing. *2021 Data Compression Conference (DCC)*, pp. 341–341, Mar 2021. doi: 10.1109/dcc50243.2021.00051.

Kadhim, D. J., Mosleh, M. F., and Abed, F. A. Exploring text data compression: A comparative study of adaptive huffman and lzw approaches. *BIO Web of Conferences*, 97:00035, 2024. doi: 10.1051/bioconf/20249700035.

Morris, J., Kuleshov, V., Shmatikov, V., and Rush, A. Text embeddings reveal (almost) as much as text. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460, 2023. doi: 10.18653/v1/2023.emnlp-main.765.

Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.-W., and et al. Large dual encoders are generalizable retrievers. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.669.

Palaniappan, V. and Latifi, S. Lossy text compression techniques. *ICCS 2007*, pp. 205–210, 2007. doi: 10.1007/978-1-84628-992-7_28.

Witten, I. H., Bell, T. C., Moffat, A., Nevill-Manning, C. G., Smith, T. C., and Thimbleby, H. Semantic and generative models for lossy text compression. *The Computer Journal*, 37(2):83–87, Jan 1994. doi: 10.1093/comjnl/37.2.83.

Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977. doi: 10.1109/tit.1977.1055714.