Lossy Text Compressor Using Quantization

Link to my Github: LINK

Giuseppe Carnicella

Abstract

In this article we are going to analyze the problem of Lossy Text Compression, a very complex challenge, often associated to images or videos instead of texts. We developed an approach that exploits pre-trained embedding models to represent text in a vector space, followed by a quantization process to further compress these embeddings. The quality of the reconstructed text was evaluated through standard metrics such as cosine similarity, BLEU score and ROUGE metrics, experimenting with different levels of quantization. The results show that despite the significant reduction in data size, the reconstructed text retains a good part of the original meaning and style, showing a promising trade-off between compression and semantic accuracy.

1. Introduction

Data compression is a fundamental technique used to reduce the size of data while mantaining its information content. We have 2 types of data compressions:

- **Lossy compression**: that sacrifice parts of the informations for a bigger reduction (commonly used for images, audio and videos)
- **Lossless compression**: that reduce the data dimensions without lose any information, but leading to a minor reduction. However, this technique is rarely applied to text, where even a small loss of information can severely affect the comprehensibility and consistency of the content.

In this project, we explore an innovative idea: ***the application of lossy compression to text***. We define lossy text compression as a process that reduces the size of text data while preserving its meaning and style, without resorting to summarization techniques. The main objective of this project is to evaluate the effectiveness of different levels of quantization in reducing the size of the embeddings while maintaining acceptable quality of the reconstructed text. To do this, we used a pre-trained embedding model called gtr-t5 (Ni et al., 2021) and then we experimented with various levels of compression. After that the vectors are dequantized and converted back to text through a reconstruction process and finally we evaluated the reconstruction quality through standard metrics such as cosine similarity, BLEU score, and ROUGE score. The results obtained allow us to analyze the trade-off between compression and reconstruction quality, highlighting potential future applications of lossy compression in text context.

2. Related work

Text compression is an established area of research, with classic techniques such as lossless compression preserving the integrity of the original content. However, the idea of applying lossy compression to text, which is more common in domains such as audio and video, is relatively new and innovative. One of the fundamental works in lossless compression is the Lempel-Ziv (LZ) algorithm, described in (Ziv & Lempel, 1977) which forms the basis of formats such as ZIP and GZIP. These algorithms reduce the size of data without loss of information. Although our project focuses on lossy compression, these techniques provide a useful contrast for understanding the innovation introduced by lossy compression in the textual context. The main article describing lossy text compression is also explored in (Witten et al., 1994) This work suggests the use of semantic and generative models to reduce the size of text while maintaining the overall meaning, sacrificing some details to achieve higher compression. Our research expands this concept by exploring how different levels of embedding quantization affect the quality of reconstruction. A central aspect of our project is the reconstruction of text from embeddings, using the model provided by (Morris et al., 2023) This study shows that despite the compressed representation of the embeddings, it is possible to reconstruct the original text with surprising accuracy.

Email: Giuseppe Carnicella <carniella.1950329@studenti.uniroma1.it>.

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

3. Method

In this project, we explored lossy compression of text by quantizing text embeddings. The methodology adopted consists of several key steps:

1. **Textual Embedding Generation:** We used the GTR-T5-Base pre-trained model (Ni et al., 2021) for its ability to produce dense and accurate semantic representations of textual data. The process of generating embeddings is done through **tokenization** where the text is divided into tokens and **encoding** operations where the model generates an embedding for each sentence by applying a mean pooling operation on the hidden state of the last layer of the model.

2. **Quantization:** To reduce the size of embeddings, we applied a uniform quantization technique realized using a linear scale with a variable number of bits, which allows us to control the degree of compression.The quantization process includes:

   - Scaling of Embeddings: Embeddings are normalized with respect to their minimum and maximum values, mapping the values to a range between 0 and 2^{bits-1}.
   - Quantization: Normalized embeddings are then rounded to the nearest integer and converted to uint8 (to save memory space).

3. **Dequantization and text reconstruction:** The quantized embeddings were dequantized to approximate the original values and then used to reconstruct the original text using the vec2text model (Morris et al., 2023).

4. **Evaluation of the Quality of Reconstruction:** The quality of text reconstruction was assessed using three main metrics:

   - **Cosine similarity:** measures the similarity between the original (A) and reconstructed (B) embedding vectors:

   $$\frac{A \cdot B}{\|A\|\|B\|}$$

   - **BLEU Score:** evaluate the quality of the translation based on the n-gram correspondence between the reference text and the reconstructed text:

   $$\exp\left(\sum_{i=1}^{n} \frac{Pn}{N}\right)$$

   where Pn is the precision of the n-grams and N is the number of orders.

   - **Rouge Score:** measures the overlap between word sequences in the original and reconstructed text:

   $$ROUGE-1 = \frac{unigrams - in - common}{unigrams - in - the - text}$$

   $$ROUGE-2 = \frac{bigrams - in - common}{bigrams - in - the - text}$$

   $$ROUGE-L = \frac{len(longest - shared - seq)}{len(seq - in - the - text)}$$

4. Results

Text reconstruction quality results were obtained for different quantization levels (2, 4, 6, 8, 10 and 12 bits) and are displayed in the Table 1. Tests were run on two settings: a medium-length text and a set of 50 sentences that showed extremely positive and consistent results. From the results we can see that the best levels of quantization for maintaining reconstructed text quality are 6 and 8 bits. At these levels, both Cosine Similarity, BLEU and ROUGE scores are significantly higher than at lower and higher levels, suggesting that too much or too little quantization severely impairs reconstruction quality.

Table 1. Results of the evaluation of text reconstruction quality for different levels of quantization

Bit	Cosine	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
2	0.768	0.033	0.406	0.177	0.365
4	0.962	0.506	0.780	0.632	0.745
6	0.991	0.882	0.939	0.891	0.923
8	0.994	0.908	0.940	0.912	0.931
10	0.096	1.97e-232	0.025	0.000	0.025
12	0.106	1.66e-232	0.025	0.000	0.025

5. Discussions and Conclusions

The project successfully explored the application of lossy compression on text, a relatively new and under-explored field. Results indicate that it is possible to reduce the size of text embeddings through quantization while maintaining a good level of reconstructed text quality, especially at intermediate levels of quantization (6-8 bits). However, excessive or poor quantization leads to significant loss of information, as demonstrated by the results for 2 and 12 bits. This study paves the way for further research in the area of lossy compression for text, with potential applications in resource-limited environments where data size reduction is crucial. In the future, it will be interesting to explore other techniques for quantizing embeddings and compare them with those used in this project, as well as to apply these methods to different types of texts and languages to test their effectiveness and generalizability.

References

Morris, J. X., Kuleshov, V., Shmatikov, V., and Rush, A. M. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*, 2023.

Ni, J., Qu, C., Lu, J., Dai, Z., Ábrego, G. H., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M.-W., et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

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, 01 1994. ISSN 0010-4620. doi: 10.1093/comjnl/37.2.83. URL https://doi.org/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, 1977.