Generative Language Autoencoders Are Lossless Embedding Models

Yifan Zhang et al.

[1]IIIS, Tsinghua University
[2]Shanghai Qizhi Institute
zhangyif21@mails.tsinghua.edu.cn
February 5th, 2024

Abstract

Generative Language Autoencoders (GLAE) are lossless embedding models.

1 Introduction

Generative Language Autoencoder: GLAE.

Use SFT for simple implementation:

Input:

Encode this text into 1-word length embedding: $< text > [text] < /text >$

Response:

$$< embedding > [embedding] < /embedding >$$

Input:

Decode this embedding into text: $< embedding > [embedding] < /embedding >$

Response:

$$< text > [text] < /text >$$

Figure 1: SFT data for training Generative Language Autoencoder.

One can use Total Coding Rate (TCR) (Ma et al., 2007; Li et al., 2022) uniformity loss:

$$\mathcal{L}_{1,\text{TCR}} = \text{TCR}(\mathbf{embedding}_{j\in\{1,\cdots,B\}}). \tag{1}$$

One can also use the repulsive part of InfoNCE contrastive loss (Chen et al., 2020):

$$\mathcal{L}_{1,\text{contrastive}} = -\log \frac{1}{\sum_{j=1}^{B} \exp\left(\text{sim}\left(\mathbf{embedding}_1, \mathbf{embedding}_j\right)/\tau\right)}. \tag{2}$$

$$\mathcal{L}_2 = -\frac{1}{N}\sum_{i=1}^{N} \log P_i(\text{text}_i|\mathbf{embedding}, \text{text}_{<i}). \tag{3}$$

Preprint. Under review.

Alternative optimization or simply add these two losses:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2. \tag{4}$$

We use the -2 layer representation as our embedding.

2 Related Work

Transformer (Vaswani et al., 2017).

3 Preliminaries

3.1 Lossless compression.

4 Language AutoEncoder

5 Iterative Embedding: Language Modeling with Infinite Context Length

Context:

$< system >$ You are a helpful assistant. $< /system >$

$< embedding >$ $\text{Embedding-of}($[previous context[:-2048]]$)$ $< /embedding >$

$< text >$ [previous context[-2048:-1]]

Figure 2: Language Modeling with Infinite Context Length.

$$\text{Embedding-of}(\text{[previous context[:-2048]]}) = \textbf{Embed}(\text{Embedding-of}(\text{[previous context[:-4096]]}) + \text{[previous context[:2048]]}) \tag{5}$$

Input:

Encode these embedding and text into 1-word length embedding:

$< embedding > [embedding\_1] < /embedding >< text > [text] < /text >$

Response:

$< embedding > [embedding\_2] < /embedding >$

Figure 3: SFT data for training capabilities.

6 Experiments

7 Conclusion

TODO

References

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020. 1

Li, Z., Chen, Y., LeCun, Y., and Sommer, F. T. Neural manifold clustering and embedding. *arXiv preprint arXiv:2201.10000*, 2022. 1

Ma, Y., Derksen, H., Hong, W., and Wright, J. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29 (9):1546–1562, 2007. 1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2