

1. In the first half of the lab, you produced t-SNE plots of BERT embeddings. You plotted these embeddings in different colors depending on which category of text they came from. How do you interpret the results? What difference(s) did you observe between the plot in Problem 1 and that in Problem 2? Summarize your observations in a few sentences.

In problem1, we visualized t-SNE plots of BERT embeddings for individual words. The clusters of different colors showed a clear separation of based the contextual meanings given to cluster together showing the contextualized property of BERT embeddings.

I problem2, we visualized t-SNE plots of BERT embeddings for entire sentences using the [CLS] token embeddings. The plot showed a clear clustering of sentence representations. The separation between clusters is less distinct compared to individual words which suggest that embeddings capture a more generalized contextual meaning with each category.

2. In Task 5.2, you saw the extractive & abstractive summaries side-by-side, as well as the ROUGE-2 scores computed for them. Which method obtained the higher ROUGE-2 score in your testing? Which method produced the "better" summaries in your opinion?

We obtained a higher ROUGE-2 score for the extractive summarization method (0.1706) compared to the abstractive method (0.0312). Since the ROUGE-2 score is higher for the extractive method, it is considered better. This method strives to preserve the words from existing sentences in the source, ensuring the most relevant wordings and preserving the context of the input. Whereas abstractive summarization attempts to generate the words that might not be present in the source text. It has the potential to produce more concise and human-like summaries but can be challenging, and the generated text might sometimes deviate from the original content.

3. In Task 5.3, you ran the text generation with different temperature values. What happens when the "temperature" is close to zero? What happens for higher temperature values (> 1)? At which temperature setting did the model generate the "best" summaries, in your opinion?

We have experimented with different values of temperature and observed the results of the model. At low temperature values close to 0, the results are more concise and at higher temperatures above 1, the results are more random and diverse. My best setting for temperature is 0.1.