Report - Problem 9.60

Juliana Carvalho de Souza

July 2024

1 Introduction

The 9.60 Problem, from the category ”Studying Learned Features in Language Models”, proposes making dimensionality reduction over neuron activations across a bunch of text, and then check how interpretable the resulting directions are.

In other words, dimensionality reduction techniques (such as PCA, SVD) can be applied to the activations of neurons across a bunch of texts to explore the interpretability of the resulting directions or patterns.

2 Solution

A solution would be to check how interpretable the results are on each layer, to see the evolution of the model classification.

3 Work

As my object of study, I selected BERT, ”encoder-only” transformer architecture, as a pre-trained model to be analysed.

My solution to this problem is based on separating positive (1) and negative (0) comments in the vector space – the better the model, the better the separation. Then, we would visualize using a dimension reduction (PCA) of the vectors in 2 dimensions.

Also, I plotted the [CLS] of the text through the layers of the model. It’s possible to see that the separation gets greater as the text passes through more layers.

In my method, I used real data from IBMD movie reviews’ compared with Toy data samples extracted from Amazon reviews.

Use PCA (after scaling the vectors) to reduce dimension and do the 2D plot Compared the accuracy through the layers - evaluated separability with logistic regression and cross-validation.

As a result, we could see good direction separation in the plot. The experiments suggest that BERT is able to differentiate distinct concepts and that this differentiation becomes more sophisticated through its layers. However, the performance did not increase much in the later layers – probably due to overfitting. Also, the Toy data did not worked well. One hypothesis is that, as raw data was not cleaned, it can contain grammar mistakes or other problems the model is not capable of capturing the separation at the last layers.