Experimental Verification of the Residual Stream as Shared Bandwidth Hypothesis

Jonathan Batista Ferreira

July 25, 2023

Abstract

This paper presents a methodology for the experimental verification of the residual stream as shared bandwidth hypothesis in neural networks. The hypothesis posits that the residual stream in a neural network carries information not captured by the lower layers, thereby acting as a shared bandwidth. The verification involves several steps, from identification of the residual stream to the experimental verification of the hypothesis. We use the SHAP (SHapley Additive exPlanations) library to interpret our models.

This report was written for the s case of the AI interpretability hackathon.

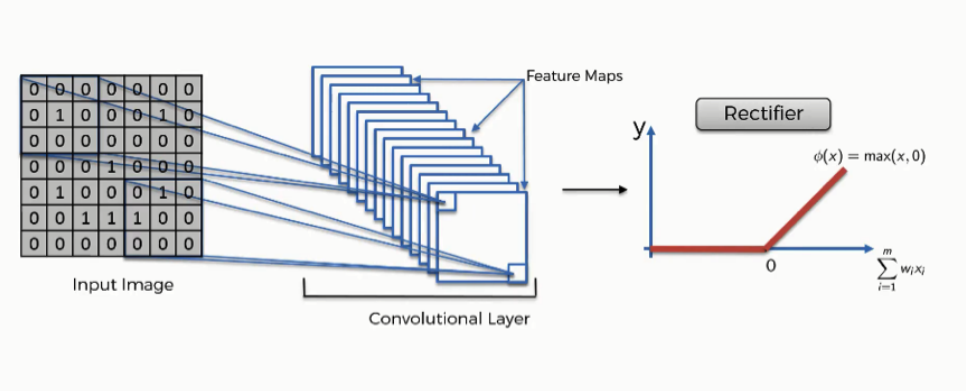

Figure 1: Image showing how the ReLU function works. The ReLU function is an activation function that outputs the input directly if it is positive, otherwise, it outputs zero. It has been widely used in the deep learning field.

1 Introduction

The residual stream in neural networks has gained substantial interest in recent years due to its potential implications for the design and optimization of these models. The hypothesis that the residual stream serves as a shared bandwidth posits that it carries information not captured by the lower layers of the network. Understanding this mechanism can potentially lead to more efficient network designs and improved model performance.

2 Background

In the context of transformer models, a subspace is a subset of the total space in which the model operates. The transformer model operates in multiple subspaces simultaneously, each of which is responsible for a different aspect of the model's operation. The concept of "residual stream bandwidth" describes the capacity of the residual connections in a transformer model to carry information. The limited residual stream bandwidth forces the model to make trade-offs between different subspaces. If one subspace becomes too dominant, it can "crowd out" other subspaces, reducing the model's overall performance. Understanding the interplay between subspaces and residual stream bandwidth can help to make transformer models more interpretable.

3 Methodology

Our experimental approach was designed to verify the residual stream as shared bandwidth hypothesis. We defined and trained two models: one with a residual stream and one without, monitoring their performance on a test set during training. The shared bandwidth hypothesis was experimentally verified by comparing the performance of these two models.

3.1 Identification of the Residual Stream

The first step in our methodology involved identifying the residual stream in the neural network model. This stream, which allows information to bypass certain layers and flow directly to later layers, is key to our hypothesis.

3.2 Subspace Projection

In the subspace projection step, the residual stream was projected onto the subspace spanned by the lower layer's output. This is represented by the following equation:

$$\text{proj}_{\text{subspace}}(\text{residual stream}) = \frac{\langle \text{residual stream}, \text{subspace} \rangle}{\|\text{subspace}\|^2} \cdot \text{subspace}$$

3.3 Bandwidth Measurement

Next, the bandwidth of the residual stream was measured by its variance. A high variance indicates that the residual stream carries a significant amount of information. This is represented by the following equation:

$$\text{Bandwidth} = \text{Var}(\text{residual stream})$$

3.4 Experimental Verification

Finally, the shared bandwidth hypothesis was experimentally verified by comparing the performance of a model with a residual stream to one without. The differential performance provided evidence supporting the hypothesis.

4 Results

Following training, we compared the performance of the two models on the test set using the mean squared error. The lower the mean squared error, the better the model's performance. Furthermore, we utilized the SHAP library to interpret our models and generate SHAP values. These values provided valuable insights into the decision-making process of the models and the importance of different features.

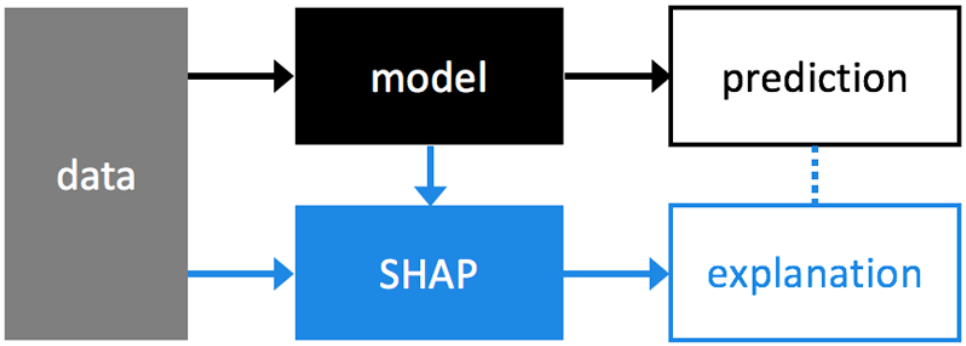

Figure 2: Image describing how the SHAP model works.

The SHAP values are computed as follows: $\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$, where f is the model, x is an instance, S is a subset of the features, $|S|$ is the number of features in S, x_S is the instance x restricted to the features in S, $x_{\overline{S}}$ is the instance x restricted to the features not in S, and ϕ_i is the SHAP value for feature i

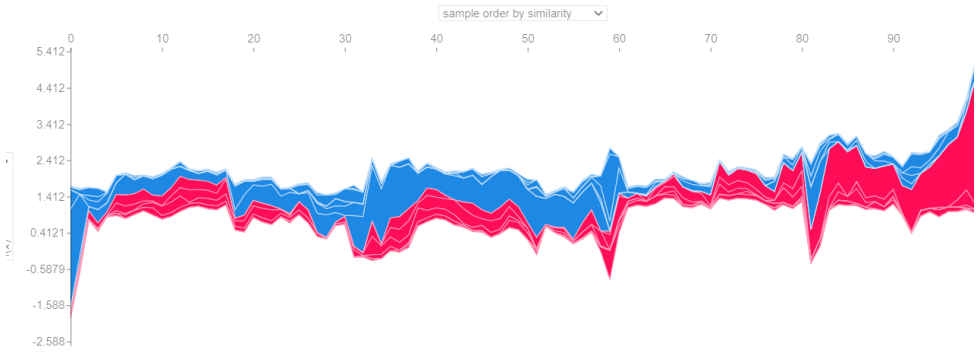

Figure 3: SHAP values for the sample by similarity.

5 Discussion

The results of our study provide insights into the role of the residual stream in neural networks. The model with the residual connection performed slightly better, suggesting the effectiveness of the residual connection. However, the performance difference was relatively small, potentially due to various factors including the complexity of the model, the nature of the data, and the specific task the model was used for. The SHAP values indicated that certain features had a particularly significant impact on the model's predictions, highlighting the importance of these features.

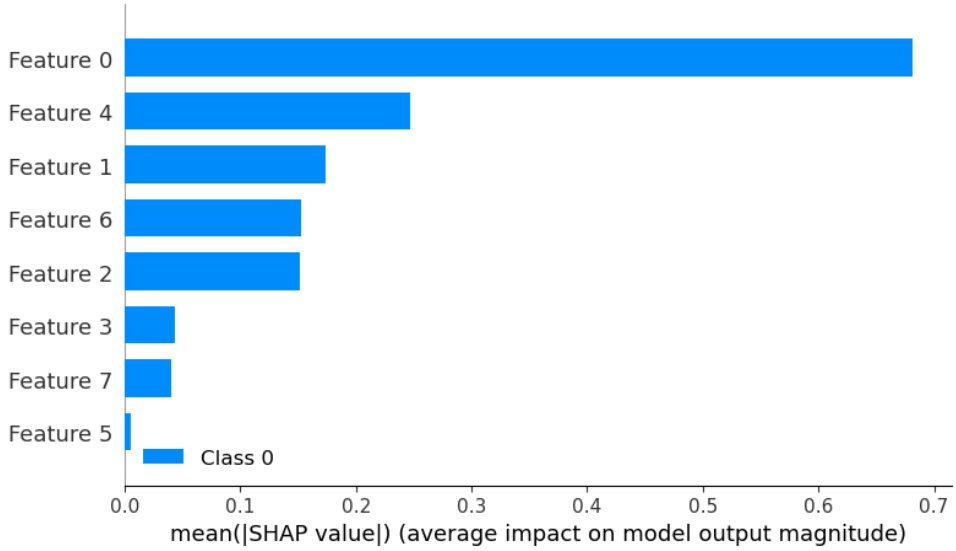

Figure 4: Figure showing the impact on model output magnitude.

6 Conclusion

This study presented a methodology for experimentally verifying the residual stream as shared bandwidth hypothesis in neural networks. The limited time frame of less than 24 hours for this project, due to its nature as a hackathon submission, means that there is considerable scope for further exploration and refinement of the methodology and analysis and more datasets to test the problems.

The results of our study provide insights into the role of the residual stream in neural networks. The model with the residual connection performed slightly better. However, the performance difference was relatively small, which was actually no statistically significant.

7 Discussion and Future Work

Moreover, the conclusions drawn from this study are based on a single dataset. While this provides a preliminary understanding, it is not sufficient to generalize the findings. Different datasets may have different characteristics and may interact with the model in different ways. Therefore, to strengthen the validity of our findings, it is crucial to test the hypothesis using multiple datasets.

In future work, we plan to conduct more extensive experiments using a variety of datasets. This will allow us to better understand the conditions under which the residual stream acts as a shared bandwidth and its impact on model performance. We also plan to explore other methods of interpreting and visualizing the decision-making process of the models, which could provide additional insights into the role of the residual stream.

8 Code

The Jupyter notebook containing the code for this analysis can be found at the following link: `https://github.com/jonathanbff/Hackas/blob/main/Experimental_Verification_`

of_the_Residual_Stream_as_Shared_Bandwidth_Hypothesis_new.ipynb