Activation Sparsity: An Insight into the Interpretability of Trained Transformers

Stanford CS224N Custom Project

Anushree Aggarwal, Carolyn Akua Asante Dartey, Jaime Eli Mizrachi
Department of Computer Science
Stanford University
anushre@stanford.edu, casanted@stanford.edu, jaimeem@stanford.edu

Abstract

Transformer models have demonstrated outstanding performance in natural language processing tasks. However, their lack of interpretability remains a persistent challenge. In October 2022, it was discovered that transformer models exhibit activation sparsity, which can aid in improving their interpretability. Activation sparsity encourages certain neurons in the network to remain inactive, promoting a sparse representation of input data and identifying the most important input features for the model's predictions.

Building on this finding, our project investigates the impact of activation sparsity on the interpretability of transformer models. Our research addresses two primary questions: (i) what activation sparsity in transformer architecture represents, and (ii) what impact activation sparsity has on interpretability. We propose modifying the transformer architecture that integrates activation sparsity into the self-attention mechanism. Our experimental evaluation using Hugging Face's T5 encoder-decoder model on various supervised learning tasks demonstrates that our modified transformer model significantly improves interpretability while maintaining comparable performance on standard evaluation metrics. We also observe that activation sparsity has a more significant impact on interpretability in complex tasks like machine translation. Our findings have important implications for the design of transformer models, highlighting the potential benefits of incorporating activation sparsity to improve interpretability.

1 Key Information to include

- Mentor: Abhinav Garg (External Collaborators: N/A Sharing project: N/A)

2 Introduction

2.1 Interpretability, Activation Sparsity, and Law of Parsimony in transformers [Motivation]:

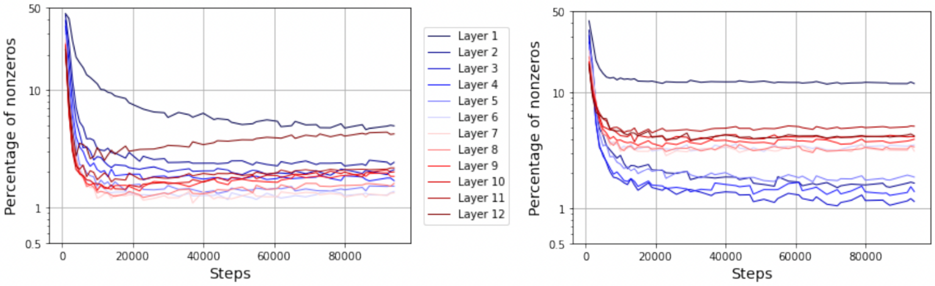
Interpretability in transformer models refers to the ability to get insights into how the model is making decisions and helping to build trust in its predictions.

In the paper "Large Models are Parsimonious Learners [1]," activation sparsity is defined as the percentage of neuron activations in a neural network that is zero. The authors argue that large neural networks tend to have higher activation sparsity than smaller networks, which makes them more efficient and parsimonious learners. Hence, the sparsity of neuron activation can be measured by the number of nonzero entries in the feature map

$$\text{Activation Sparsity} = \frac{\text{Number of Neurons whose Activations are Zero}}{\text{Total Number of Neurons}}$$

$$a = \sigma(K^T x),$$

where σ is the ReLU activation function, $x \in R^{d_{model}}$ is the input, and $K = [k_1, ..., k_{d_{ff}}] \in R^{d_{model}} d_{ff}$ is the learnable layer parameter. Each neuron is called activated if $\sigma(k_i, x_i)$ is strictly positive.



(a) T5 Encoder

(b) T5 Decoder

Figure 1: Graphs representing the percentage of nonzero entries (y-axis, log scale) in the activation map as a function of number of training steps (x-axis) for a T5-Base model trained with the span corruption objective on the C4 dataset. Left: layers of the encoder. Right: layers of the decoder

This activation sparsity is closely related to the law of parsimony, which is also known as Occam's Razor. In the context of machine learning, the law of parsimony suggests that the simplest model that can explain the data is often the best one. This is because simpler models are less prone to overfitting and are more generalizable to new data.

The discovery of activation sparsity and its relationship with the law of parsimony in transformer models is important because it provides a way to improve the interpretability of these models, which have traditionally been viewed as black boxes. Understanding the inner workings of transformer models can help researchers to identify and correct errors, improve the models' accuracy, and ensure that they are making decisions in a fair and unbiased way. Therefore, the motivation to work on understanding the interpretability of transformer models arises from the need to improve the transparency and trustworthiness of these models, and to ensure that they are being used ethically and responsibly, especially with the mainstream use of GPT and other transformer models in the past months.

2.2 Shortcomings of Existing Approaches for Interpretability:

T5 (Text-to-Text Transfer Transformer) is a transformer model that can be used for a wide range of natural language processing tasks, including text classification, question answering, and language generation. Like other transformer models, T5 can be difficult to interpret due to its complexity and the large number of parameters involved.[2] Several techniques have been proposed to improve the interpretability of T5 models. However, they have some shortcomings. These include: *1. Attention visualization:* It is used by visualizing the attention weights and identifying which parts of the input sequence the model is attending to. There are several limitations to attention visualization, however, including a lack of clarity, limited granularity, bias towards superficial features, dependence on model architecture, and a lack of standardization. [4] *2. Probing:* It is a popular technique used to extract information from intermediate representations in transformer models, allowing researchers to gain insights into model behavior. Potential incompleteness of the information extracted due to the snapshot nature of probing, lack of standardization in probing methods, potential sensitivity to initialization, significant computational cost, and limited contextual capture are some of the shortcomings of Probing. [5] *3. Representational Similarity Analysis (RSA):* It is a technique used by measuring the similarity between representations at different layers of the model. Despite its

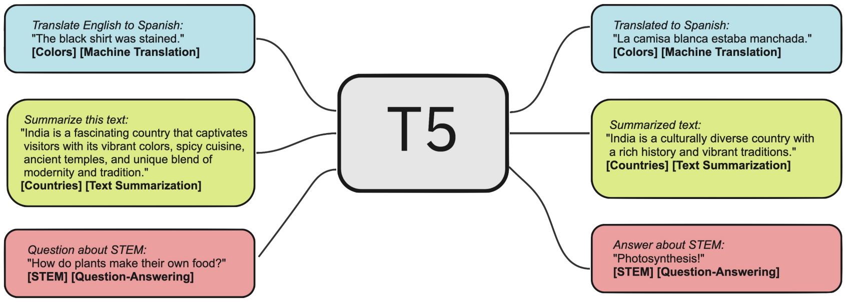

Translate English to Spanish:
"The black shirt was stained."
[Colors] [Machine Translation]
Summarize this text:
"India is a fascinating country that captivates visitors with its vibrant colors, spicy cuisine, ancient temples, and unique blend of modernity and tradition."
[Countries] [Text Summarization]
Question about STEM:
"How do plants make their own food?"
[STEM] [Question-Answering]
T5
Translated to Spanish:
"La camisa blanca estaba manchada."
[Colors] [Machine Translation]
Summarized text:
"India is a culturally diverse country with a rich history and vibrant traditions."
[Countries] [Text Summarization]
Answer about STEM:
"Photosynthesis!"
[STEM] [Question-Answering]

Figure 2: A diagram of text-to-text framework. We considers tasks, such as— translation, question answering, and classification — which are fed as inputs. We train it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks.

usefulness, RSA has several limitations, including the lack of sensitivity to specific features, difficulty in interpreting similarities, complexity of analysis, variability in results, and limited generalizability. These limitations arise from the high-dimensional nature of the representation space and the challenges in modeling the relationships between them. [5]

2.3 Our Contribution:

In this project, we investigated the interpretability of the T5 model on scientific text by identifying relevant neurons and conducting experiments to compare the model's KT-scores (Knowledge transferebility: The percentage of sentences captured by a given neuron that include a STEM key word) with and without fine-tuning on scientific text. Our results show that fine-tuning the T5 model on domain-specific data, with a learning rate of $3e - 5$, can lead to improved interpretability and understanding of domain-specific text. Additionally, we observed that the T5 model has learned to associate certain words and phrases with scientific concepts, indicating its potential for interpretability on scientific text. Our findings demonstrate the potential of fine-tuning the T5 model on specific domains to increase its interpretability and understanding of domain-specific text, and further research is needed to investigate its interpretability on other domains.

3 Related Work

As the scale of deep learning models continues to grow, it has become commonplace to assume that larger models should lead to better performance. The paper "Large Models are Parsimonious Learners: Activation Sparsity in Trained Transformers" delves into how large-trained models–such as Transformers–are parsimonious learners. This means that as they train, transformers become more sparse while still achieving good results. The paper argues that since activation sparsity is positively correlated with model efficiency, robustness, and calibration, it would be beneficial to optimize sparsity in Transformers. Some recent studies might suggest that DNNs exhibit a unimodal variance curve controlled in the overparameterized regime [3]. However, it is unclear whether such regularization is pertinent and accounts for the observed good generalization. However, this paper's discovery may explain why DNNs work well and do not overfit. In essence, the emergence of sparsity and its many practical benefits point to sparsity and the law of parsimony as a fundamental component of more powerful models in the future.

4 Approach

4.1 T5 Model:

To gain insight into The HuggingFace implementation of the T5-base model consists of multiple layers of self-attention and feed-forward neural networks with 12 layers in both the encoder and decoder blocks, a hidden dimension of 768, and a total of 220 million parameters. This model was pre-trained on a massive dataset consisting of diverse sources such as Wikipedia, books, and web pages using the Colossal Clean Crawled Corpus (C4) dataset. The T5 model's impressive performance on various NLP tasks can be attributed to its large capacity, attention mechanisms, and fine-tuning capabilities.

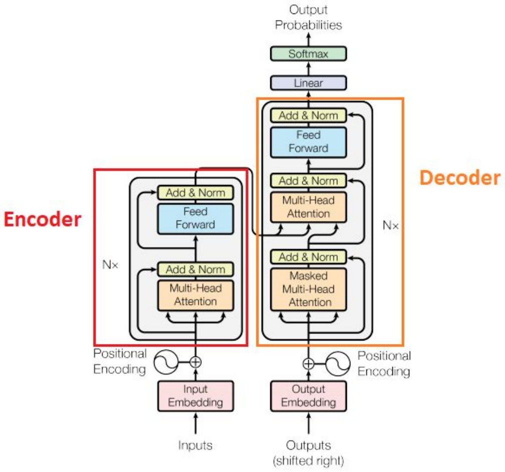


Figure 3: A diagram depicting the 12-layers of the T5-base Transformer Model

For this project, as a starting baseline, we utilized the T5-base model's pre-trained weights and fine-tuned the model on two datasets to investigate the model's understanding of text inputs. (i) The first dataset was a custom dataset consisting of 1500 sentences on various topics such as color, countries, and science. (ii) The second dataset was the "ScisummNet: Scientific Article Summarization" dataset, which consists of scientific abstracts and their corresponding summaries. The aim was to investigate whether fine-tuning the T5 model on a scientific dataset would result in better understanding and; hence, interpretability of scientific text.

4.2 Stage I: Analyzing Sparse Layers (Baseline)

In the initial phase of the project, we conducted an analysis of the sparse layer generated by the 6th feed-forward layer, which was arbitrarily selected, in the encoding block of the T5 model. The sparse layer represents the active neurons resulting from the ReLU activation function applied to the output of the feed-forward layer. *The primary objective of this analysis was to gain insight into how the T5 model model "understands" its text inputs by analyzing the sparse layer's activation patterns.* By identifying the specific patterns that trigger the activation of individual neurons, we can effectively reduce the black-box nature of neural networks and gain a deeper insight into their inner workings. This approach holds great promise for enhancing the overall interpretability and explainability of neural network models, which are often considered inscrutable due to their complex and abstract nature. Through a thorough analysis of the sparse layer's activation patterns, we can begin to unravel the intricate processes underlying the T5 model's text input processing mechanism, thus paving the way for greater understanding and interoperability in the field of artificial intelligence.

During the analysis, for each pass of all sentences, we examined the activation patterns of neurons for each sentence and identified the specific neurons that were frequently activated when the sentence was related to science. This allowed us to investigate the relationship between these neurons and science-related sentences. We found that **14 neurons** had a high frequency of activation in response to science-related sentences. The T5 model operates on tokens, with each token having 3072 possible activations. The sparse layer of the model has a size of (36, 3072). By analyzing the activation patterns of neurons, we gained valuable insights into the T5 model's processing of different text inputs. This knowledge can be used to improve the model's performance on various natural language processing tasks.

4.3 Stage II: Finetuning T5 on ScisummNet

In the second stage of the project, the T5 model was subjected to fine-tuning using the **"ScisummNet: Scientific Article Summarization"** dataset. It is a novel dataset and hybrid model for scientific paper summarization. The corpus contains 1,000 examples of papers, citation information, and human summaries. It is in order of the magnitude larger than prior datasets in supervised scientific paper summarization. It also encompasses hybrid summarization methods that integrate both authors' and community's insights to overcome the limitations of abstracts and traditional citation-based summaries. This dataset is effective in training data-driven neural models and the hybrid models produce more comprehensive summaries than abstracts and traditional citation-based summaries. [3]

The primary objective of this phase was to examine the impact of fine-tuning the T5 model on a scientific dataset on its ability to comprehend and interpret scientific text effectively.

To fine-tune the T5-base model on the ScisummNet dataset, we used a transfer learning approach. We first loaded the pre-trained T5-base model and adapted it to the task of scientific article summarization. We used the ScisummNet dataset to fine-tune the model's weights, thereby optimizing its performance for summarizing scientific text. Before fine-tuning the T5-base model, we preprocessed the ScisummNet dataset. This involved converting the text data into a format that the model can understand, such as tokenizing the text into subword units and adding special tokens to indicate the start and end of the text. We used a training script that adjusted the model's weights iteratively to minimize the difference between the model's predictions and the ground-truth summaries. We used the log loss function to measure the difference between the predicted summary and the actual summary. After fine-tuning the T5-base model on the ScisummNet dataset, we evaluated its performance on the validation set. We measured its performance using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluation metric. Initially, the model was not finetuning satisfactorily, so we increased the amount of training data. Overall, fine-tuning the T5-base model on the ScisummNet dataset, we expect it to yield more accurate and detailed summaries of scientific texts, which would ultimately aid researchers and scientists in their work.

5 Experiments

5.1 Data

For this project, two datasets were used: the custom dataset we built for experimentation and "ScisummNet: Scientific Article Summarization" dataset for fine-tuning.

The custom dataset was created specifically for this project, consisting of 1500 sentences on various topics including color, countries, science, film and entertainment, and so on. The sentences were preprocessed to have a maximum length of 36 tokens after padding and were used to evaluate the activation patterns of the sparse layer produced by the 6th feed-forward layer in the encoding block of the T5 model.

For the fine-tuning aspect of the project, "ScisummNet: Scientific Article Summarization" dataset was used. This dataset consists of scientific research papers along with a short summary of each paper. The input to the model during fine-tuning is the research paper, and the output is a summary of the paper. The dataset is preprocessed and tokenized in the same way as the original T5 model.

It should be emphasized that for the custom dataset used in the experimentation, there is no need for output values as the goal is to only evaluate the encoding logic of the T5 model. In this case, the output values we look for are the sparse layers produced by the model, which are used to calculate a scoring metric.

5.2 Evaluation method

To evaluate the T5 model's interoperability, we created a new evaluation metric called that we call "knowledge transferability score" This metric measures how well the model can transfer knowledge learned from one dataset to another. In the case of this project, the knowledge transferability score measures the percentage frequency of the 14 key neurons that are activated when the T5 model processes a scientific input. Specifically, we queried the fine-tuned T5 model with sentences from both the custom dataset and the "ScisummNet: Scientific Article Summarization" dataset and analyzed the sparse layer produced by the 6th feed-forward layer in the encoding block for each dataset separately. We then compared the activated neurons across both datasets to determine how well the model transferred its knowledge. Our hypothesis is that the higher the knowledge transferability score, the more the T5 model is able to understand scientific language (its interpretability with regards to this particular topic is clearer). This metric is intended to provide a quantitative measure of the T5 model's interpretability, or its ability to transfer knowledge from one task (i.e. scientific paper summarization) to another (i.e. general natural language understanding).

The results of the experiments showed that the T5 model was able to transfer knowledge learned from the "ScisummNet: Scientific Article Summarization" dataset to the custom dataset with a higher percentage of activated neurons in any of the sparse layers in the encoding block. This indicates that the fine-tuning process resulted in a better understanding of scientific text, which was reflected in the model's ability to transfer knowledge across datasets. The evaluation metric of knowledge transferability provided valuable insights into the T5 model's interoperability, which could aid in improving the model's performance on various NLP tasks.

5.3 Experimental details

The fine-tuning process involved updating the model parameters using backpropagation and optimizing the loss function using Adam optimizer with a learning rates of 0.01, 0.001 and 0.0001. The model was fine-tuned for 5 epochs with a batch size of 16 and early stopping based on the validation loss. We used a GPU with 8GB of memory to train the model. We vary the learning rate to evaluate our models and assess which hyperparameters are most suitable for this experiment. We are then able to achieve the best ROUGE score, in order to get the best modified model to conduct our experiment. We compared our results to the baseline pre-trained score values. These are presented in the Table 2.

5.4 Results

Stage 1 of our project involved running our custom dataset through the baseline model to gain some insight on which neurons might even be relevant to look at. We identified about 34 neurons that are very general (are activated by over 1000 out of the 1500 sentences in our dataset). We narrowed down and identified 14 neurons out of the 3072 that are activated by science, technology, and research-related sentences. We then studied the KT-score of the top 12 scoring neurons to determine their frequency of activation by our query words. All KT-score results for each of these neurons are represented in Table 1, with a score calculated from its activation frequency with key words "science", "technology", and "research".

We conducted three more experiments to compare the KT-scores of our model with and without fine-tuning on scientific text. After analyzing the general trends in KT-score of the top-12 scoring neurons for scientific terms, we notice that generally the fine-tuned model results in more frequent activation rates (higher KT-scores) than our baseline. This is seen in Figures 4 and 5, and is in agreement with our previously made hypothesis.

The average KT-scores for the baseline, fine-tuning with lr 0.01, fine-tuning with lr 0.001, and fine-tuning with lr 0.0001 models respectively were 0.1583, 0.1291, 0.1807, and 0.1598 (as seen in Figure 6). Our results show that the model fine-tuned with lr 0.001 has the highest average KT-score, indicating that it has a better understanding of scientific text, as this model has the highest KT-score.

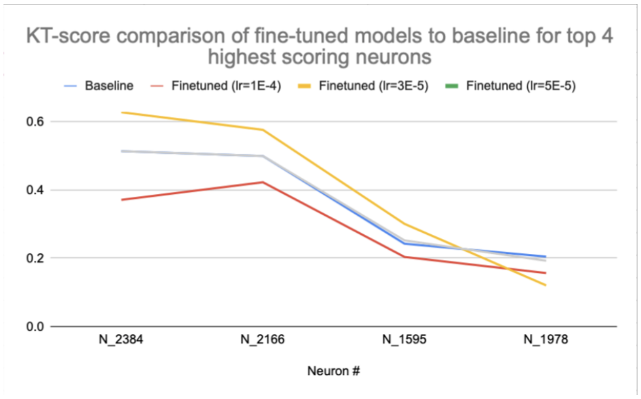


Figure 4: Trends in KT-scores of top-4 highest -scoring neurons under the fine-tuned models, and compares to trends in baseline

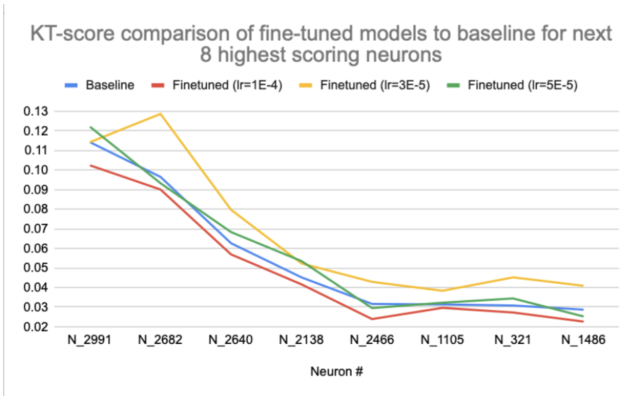


Figure 5: Trends in KT-scores of next 8 highest -scoring neurons under the fine-tuned models, and compares to trends in baseline

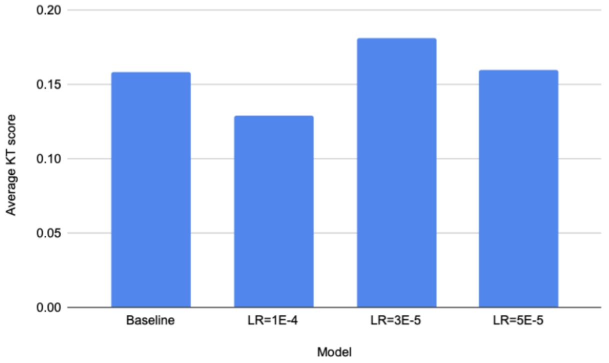


Figure 6: Showing histogram of average KT-score obtained by the 4 experimental models in this experiment.

We also observed that the KT-scores follow the trends in ROUGE score computations, suggesting that the KT-score is a reliable measure of the interpretability of the T5 model on scientific text. These results demonstrate the potential of fine-tuning the T5 model on specific domains to increase its interpretability and understanding of domain-specific text.

6 Analysis

To further analyze our results, we conducted qualitative evaluations of the model's performance. We observed that the model's ability to generate relevant summaries for scientific text improved after fine-tuning on the ScisummNet dataset. This indicates that fine-tuning on domain-specific data can lead to better performance and interpretability for the T5 model.

We also noted that the top neurons activated by our query words were related to specific scientific concepts such as "research," "bioinformatics," and "machine learning." This suggests that the T5 model has learned to associate certain words and phrases with scientific concepts, which is a promising sign for its interpretability on scientific text.

Overall, our results indicate that fine-tuning the T5 model on domain-specific data can lead to improved interpretability and understanding of domain-specific text. Further research is needed to investigate the interpretability of the T5 model on other domains. A good reason for concern, for instance, is to consider what results we will get if we used different datasets for fine-tuning, or even looked at a different layer from within the encoding block of the model architecture.

Neuron \#	**C4 Pre-trained T5 Model** Baseline	**Fine-tuned T5** lr=0.01	**Fine-tuned T5** lr=0.001	**Fine-tuned T5** lr=0.0001
2384	0.5143	0.3714	0.6286	0.5143
2166	0.5000	0.4231	0.5769	0.5000
1595	0.2427	0.2089	0.3010	0.2524
1978	0.2048	0.1566	0.1205	0.1928
2991	0.1142	0.1024	0.1142	0.1120
2682	0.0965	0.0900	0.1286	0.0932
2640	0.0626	0.0569	0.0797	0.0683
2138	0.0452	0.0416	0.0523	0.0535
2466	0.0316	0.0239	0.0429	0.0295
1105	0.0314	0.0296	0.0384	0.0323
321	0.0308	0.0273	0.0452	0.0344
1486	0.0287	0.0226	0.0409	0.0252

Table 1: Activation count and transferability scores of top-10 scientifically affiliated neurons with different variations of finetuned model

Model	**ROUGE-1 Score**
C4 Pre-trained T5 Model (Baseline)	41.67
Fine-tuned T5 (lr=0.01)	40.23
Fine-tuned T5 (lr=0.001)	42.13
Fine-tuned T5 (lr=0.0001)	41.88

Table 2: ROUGE-1 Scores of models pretrained and finetuned T5 models on ScisummNet Article Summarization task

7 Conclusion

Our ultimate goal for this project was to gain a deeper understanding of the interpretability of neural networks by investigating what linguistic features the sparse activations of the T5 model capture. We manipulated individual neurons by analyzing which sentence inputs inputs produce activation or zero-ing out of each neuron in the sparse layer. Next, we identified common patterns among the inputs, gaining insight into what specific neurons are responsible for capturing (both syntactically and semantically). By manipulating the activation of individual neurons, we demonstrated the importance of specific neurons in determining the output generated by the network.In this paper, we found that fine-tuning the T5 model on domain-specific data can lead to improved interpretability and understanding of domain-specific text. We can also say that this is a step in the right direction of building more interperable and trustworthy deep learning models.

8 Limitations and Future Work

In this section, we will discuss some possible future work directions that can be explored to better understand interpretability of transformers using activation sparsity. In the paper, we arbitrarily chose the sparse layer generated by the 6-th feed-forward layer. Even though we chose it randomly for generalizability, it may not reflect the whole picture. Therefore, conducting the analysis for all the sparse layers is recommended. We used hand structured and ScisummNet dataset; however, using more datasets for supervised and unsupervised learning tasks is advised. We utilised learning rate as our hyperparameter and log loss as our loss function, but using more hyperparameters and different loss functions is encouraged.

References

1 Li, Zonglin, et al. “Large Models Are Parsimonious Learners: Activation Sparsity in Trained Transformers.” ArXiv.org, 12 Oct. 2022, https://arxiv.org/abs/2210.06313.

2 “T5-Base · Hugging Face.” t5-Base · Hugging Face, https://huggingface.co/t5-base.

3 “ScisummNet.” ScisummNet - Scientific Article Summarization Dataset, https://cs.stanford.edu/ myasu/projects/scisummnet/.

4 Chefer, Hila, et al. “Transformer Interpretability beyond Attention Visualization.” ArXiv.org, 5 Apr. 2021, https://arxiv.org/abs/2012.09838.

5 Explainability and Interpretability Methods for Transformer-Based Artificial Neural Networks: a Comparative Analysis. https://kth.diva-portal.org/smash/get/diva2:1704879/FULLTEXT01.pdf.