Activation Sparsity:
An Insight Into Trained Transformer Interpretability

Carolyn Asante Dartey, Anushree Aggarwal, Jaime Eli Mizrachi
Department of Computer Science, Stanford University

Introduction

Problem

Transformers are used in everyday life now, but can we trust them to make high stake decisions if we don't even understand how they are making these decisions?

Background

In October 2022, activation sparsity in transformers was discovered, which can help with more interpretable models .

Research Questions

i) What does activation sparsity in transformers represent?

ii) How does it impact interpretability?

Methods

Model

We use the HuggingFace T5-base (Text-to-Text Transfer Transformer) model pre-trained on Colossal Clean Crawled Corpus (C4) dataset.

Data

(i) Custom Dataset on colors, countries, and STEM.
(ii) ScisummNet: Scientific Article Summarization

Evaulation Metric

Introduced a new evaluation metric called "Knowledge Transferability score (KT)*" to measure model's ability to transfer knowledge from one dataset to another,

Experiment and Results

- Adam optimizer with a learning rates of 0.01, 0.001 and 0.0001.
- Fine-tuned for 5 epochs
- Batch Size: 16
- Used a GPU with 8GB of memory to train the model.

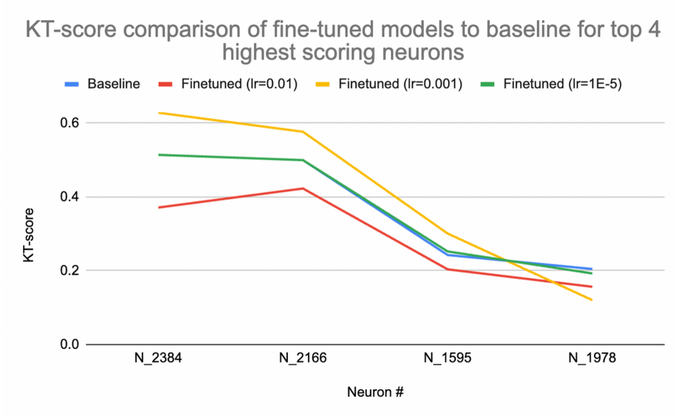
STAGE I: Analysing Sparse Layers

- Method: Analyzed the sparse layer generated by the 6th feed-forward layer in the encoding block
- Result: Potential for improving model's natural language processing performance

STAGE II: Finetuning T5-base on ScisummNet Dataset

- Method: Transfer learning method
  - Preprocessed by tokenizing the text and adding special tokens
  - Used log loss function to adjust the model's weights iteratively
- Result: Yielded more accurate and detailed scientific text summaries.

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

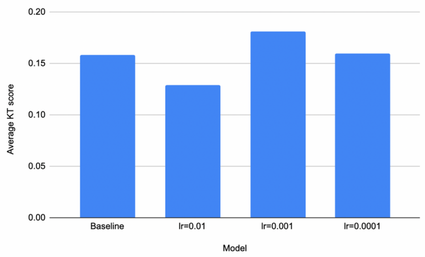


KT-score: The percentage of sentences captured by a given neuron that include a STEM key word.

Analysis

Key insights from our experimentation:

- 34 neurons in the baseline model activated by general sentences.
- 14 out of 3072 neurons consistently activated by science, tech, and research sentences.
- 35.58% of neurons weren't activated on any input
- KT-scores of top neurons:
  - To see activation by query words.
  - After fine-tuning: Increase in KT-score, suggesting a better understanding of the scientific text.
  - Follow ROUGE score trends, representing model's interpretability of scientific inputs,



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

Future Work

- Investigating all the layers
- Using more hyperparameters
- Fine-tuning on different datasets
- Trying with different loss functions

References

[1] Li, Zonglin. "Large Models Are Parsimonious Learners: Activation Sparsity In..." arXiv.org, 12 Oct. 2022, arxiv.org/abs/2210.06313.
[2] "ScisummNet." ScisummNet - Scientific Article Summarization Dataset, https://cs.stanford.edu/~myasu/projects/scisumm_net/.
[3] "T5-Base · Hugging Face." t5-Base · Hugging Face, https://huggingface.co/t5-base.