Interpretability of Transformer Circuits

A Systematic Analysis of Transformer Circuits

Hemanth Kumar Jayakumar — Avi Singhal — Rohitaa Ravikumar

Rice University

Contact Information:
Department of Computer Science,
Department of Electrical and Computer Engineering
Rice University
6100 Main St, Houston, TX 77005
Email: as278@rice.edu,hj51@rice.edu, rr71@rice.edu

Abstract

Though holding the crown for language and even vision modeling, Transformers still remain a mystery to its users, acting as a black-box solution to every problem. Several attempts have been made to analyze Transformers but have not been able to produce results that would expose its inner workings. We follow the trail of Mechanistic Interpretability, left by Nelson et. al, and try to provide meaningful interpretations of attention-only transformer models to identify a portion of the workings of a transformer. We believe this would be a starting point for a much deeper analysis of the transformer circuit laying the ground for what is to come.

Introduction

The rise of Transformers when the paper Attention is all you need was introduced brought forth an immense change in Deep Learning's ability to utilize language data. Every industry has begun to leverage its sequence and even image data pooling into a transformer model and gaining incredible, and transferable results.

While GPT2 and GPT3 have established the power of transformer models, their inner workings still remain unknown to their users and owners. There have been several attempts at interpreting transformer models but no solid proof of concept has been established for how and why transformers work. This project bridges the gap with the help of Mechanistic Interpretability. Being a new field within interpretability, we believe that this approach of attempting to unveil transformers would provide a better understanding of what the attention layers are up to, provide reasoning for the results produced by the transformer and also help to improve models.

Literature Review

C. Olah et al. attempted to reverse-engineer transformers. They started by analyzing autoregressive decoder-only transformers with 0,1,2 attention layers. They found that 1 layer attention-only transformer are ensemble of bigram and skip trigrams. The 2 layer attention-only transformer becomes more capable for in-context learning, the authors claim that it is due to the formation of induction heads, which search over the context for previous examples of the present token. If they don't find it, they attend to the first token. The authors trained their model on various datasets to establish their findings and provided mathematical reasoning for the induction heads.

Objectives

1. Train an auto-regressive 1,2 layer attention-only decoder only transformer.
2. Analyse the attention scores and outputs for each layer and understand what the attention layer has learned.

Experiment Settings

1. Datasets considered: IMDB (50,000), wikitext(1.8M)
2. Tools used: Pytorch, Kaggle, Colab
3. Hyperparameters: no. attention heads, embedding size, learning rate, optimizer, sequence length, tokenizer, epochs, size of dataset.
4. 1 and 2 Layer, Attention only autoregressive decoder only transformer with IMDB dataset with a different set of attention heads, max sequence length and embedding sizes.
5. 2 Layer Attention with MLP layers, autoregressive decoder only transformer with IMDB dataset.
6. 1 Layer Attention only autoregressive decoder only transformer with wikitext dataset with different dataset sizes, embedding size.
7. Plotted the attention scores of a few different heads of gpt2 attention layer 1.

IMDB Dataset

1, 2 layer attention-only transformer model

Both 1-layer and 2-layer attention-only transformer models were trained over the IMDB Dataset for 20 epochs for the autoregressive task of Text Generation, similar to how GPT2 operates. Over each epoch, the model did not learn anything useful. Different embedding sizes(128, 256) were attempted and did not yield any results. Changing the max length of the sequences (20, 30, 50) and the tokenizers(GPT2Tokenizer and BERTTokenizer) did not provide any outcomes we expected either. Multiple tokenizers were tested with to improve vocabulary size attempting improved model accuracy. The number of attention heads was also attempted to be tuned(4, 8). Experiments were conducted with the addition of MLP layers to the previous experimental settings in the IMDB Dataset to reach a better model training.

Wikitext Dataset

1 layer attention-only transformer model

1-layer attention-only transformer model with max len of 50, 80, embedding size 128 and 256, attention heads count of 8, trained with the half and the entire dataset for a total of 35 epochs.

Results

The experiment settings were not clearly outlined in the reference paper, this made it difficult to replicate the experiments. Below are the key difference that we could identify. Other things like: the use of pretrained embeddings, training time, list of datasets etc. were not specified.

Reference Paper Settings	Our Settings
Always used residual connections	Used residual connections only for IMDB dataset training
Used 12,32 attention heads	Used 4,8 attention heads
Used large embeddings (768)	Used smaller embeddings (128,256)
Used gpt2neo tokenizer	Used gpt2, bert tokenizer
Used huge datasets for training	Our datasets were comparatively smaller
Use of maxlength of 2048	We used maxlength up to 80

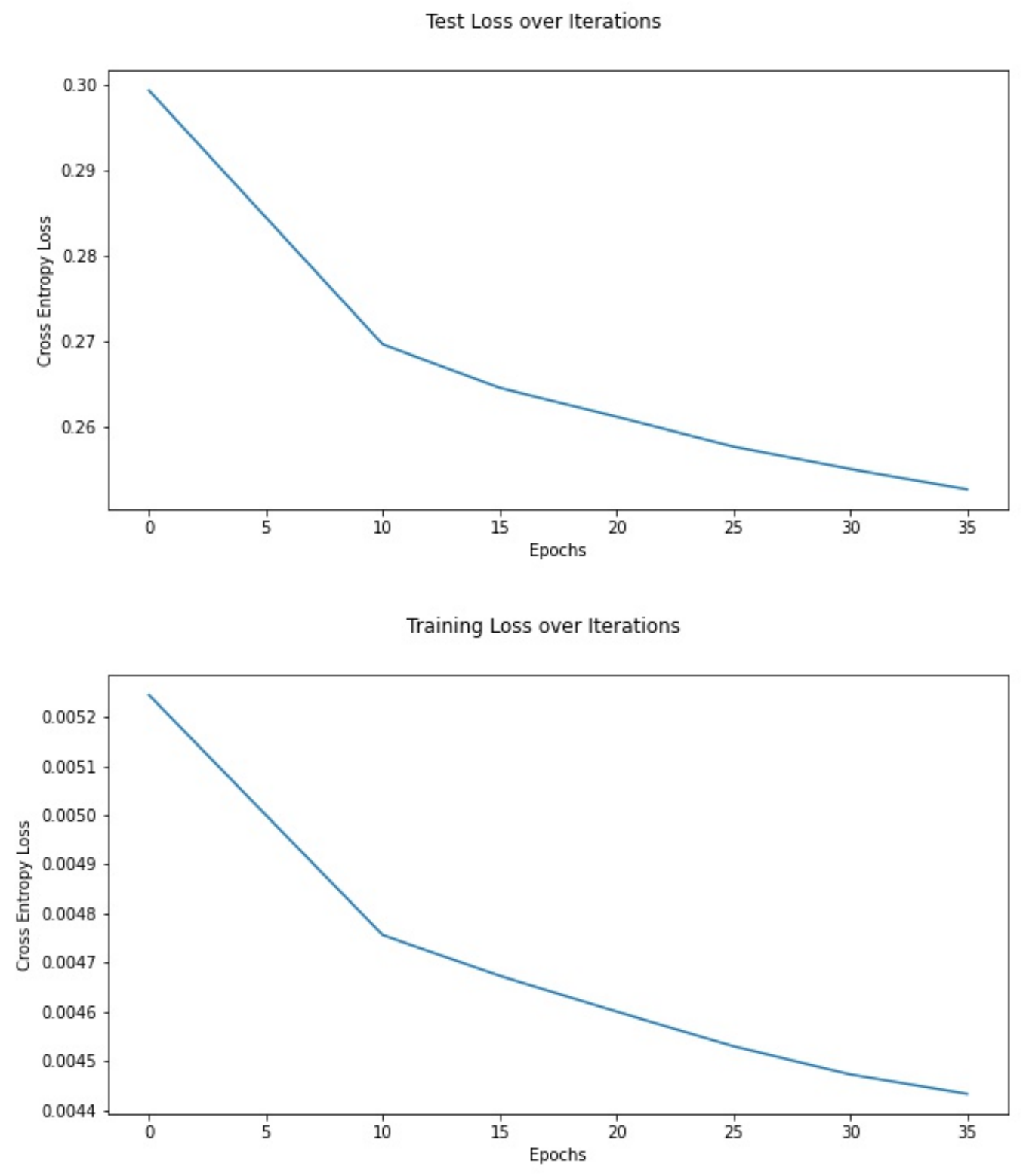

Figure 1: Loss curves

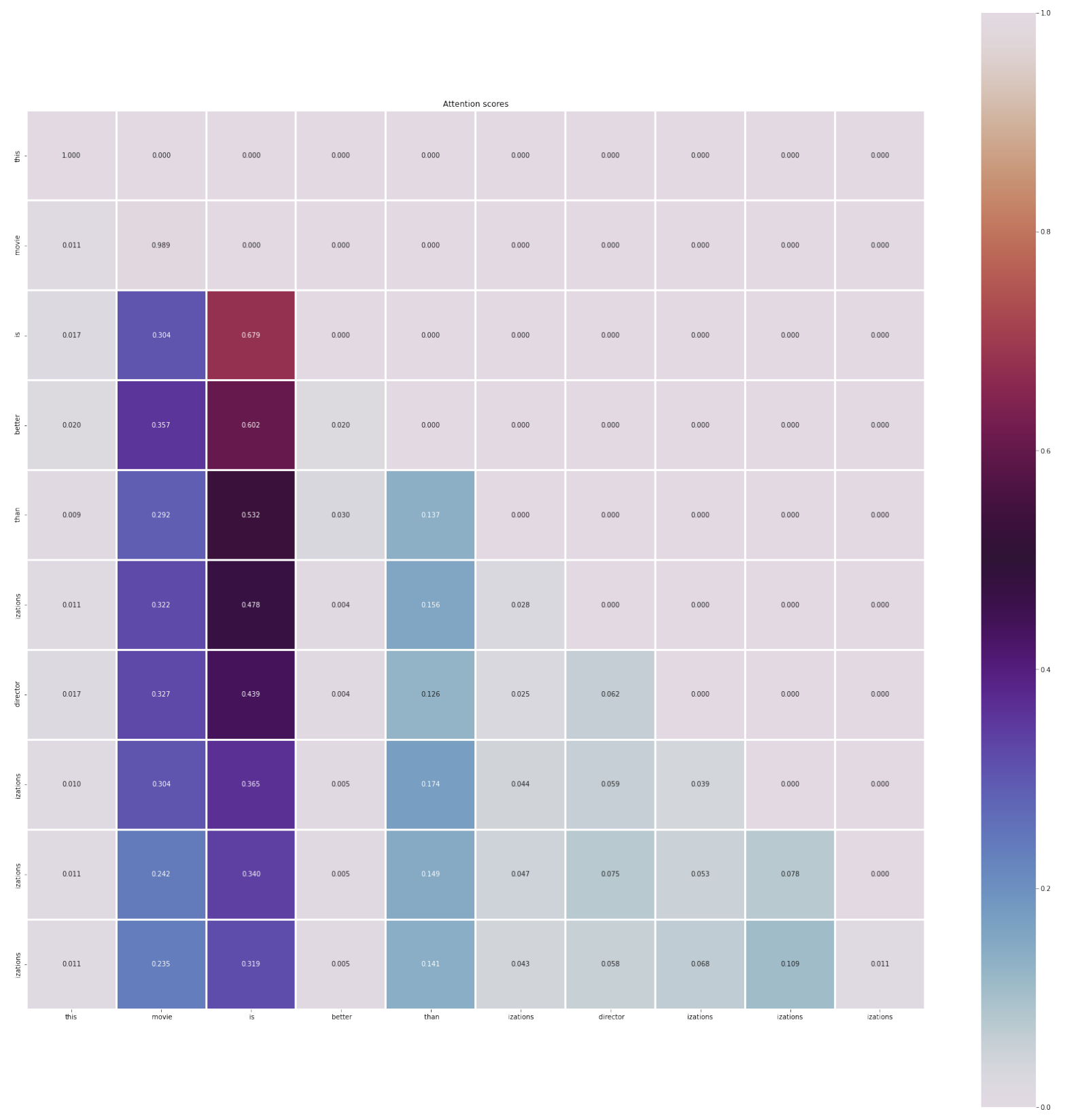

Figure 2: Attention head maps

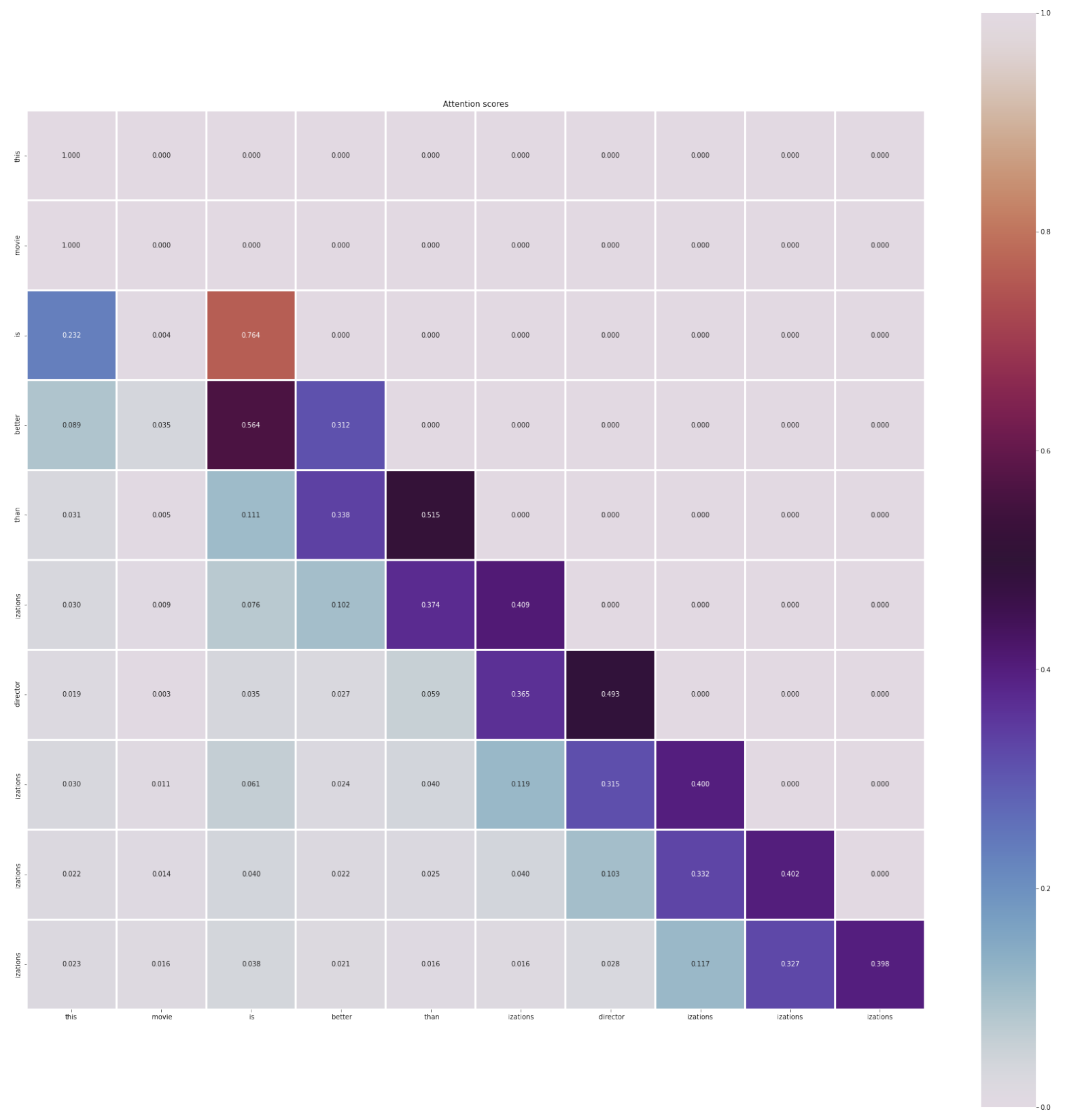

Figure 3: Attention head maps

Challenges Encountered

- Obtaining clear understanding of transformers and its mathematics.
- Creating working pipeline for decoder-only transformer.
- Training the model, using small datasets caused the model to not train well and not give meaningful attention interpretation.
- Too many hyperparameters to tune.
- Training took very long when we increased the size of the dataset. For 1 layer attention only transformer, it took 36 hours of GPU time for 45 epochs.
- We searched for a pretrained model, but even after extracting the attention layer weights, we could not directly use them as the config needed to be the same.

Conclusion

- The model is able to train which is visible from the training loss plots, but the model is not training good enough.
- IMDB dataset is too small to be used for training because it has just 50,000 rows and we have a large number of parameters to be trained (embedding layers + weight matrices Q,K,V for multiheadattention + output layer)
- The model needs to be trained on a larger dataset (like wikitext) for large number of epochs, also we think that increasing the sequence length will allow the model to learn long distance relations because the embedding also needs to be trained.
- Currently by looking at the attention scores, it seems that the attention heads are mostly focusing on the starting of the sequence, they are not really learning any contextual information because the model is not trained well.
- The use of multiple GPU for training will enable efficient training.

Forthcoming Research

We will try to look for smaller models which are pretrained and analyse the attention scores of inputs, since models without MLP layers are not available, we will try to analyze models with MLP layers.

References

[1] Nelson Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[2] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.