Interpreting GPT-2 with Emotion Classification

Neeti Desai
Northeastern University

Aditya Ratan Jannali
Northeastern University

Karen Zhang
Northeastern University

Abstract

Large Language Models have proven to be suitable for a variety of complex tasks. Despite their widespread usage, LLMs are still considered black boxes, with information encoded onto thousands to billions of parameters. Therefore, our goal is to understand the architecture and inner information representation of a LLM. To achieve this, we finetuned a GPT-2 model on an emotion classification task, and conducted three interpretability experiments to observe how the output and activations change when predicting emotion.

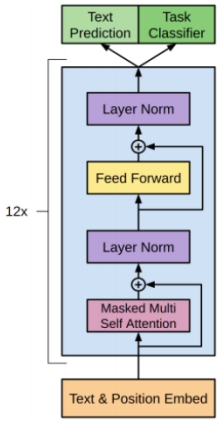

Figure 1: Architecture of GPT-2 small - [source]

1 Introduction

In this paper we explored the architecture of the GPT-2 model, to better understand how it makes predictions. We finetuned a pre-trained GPT2-small on a Kaggle Twitter Emotion Detection Dataset for an emotion classification task (kag). We then individually masked attention heads to see how the probability distribution for a sentence initially predicted with high confidence by the model changed. The intent was to observe the role of each head in the prediction, or if head(s) play any role at all. We also explored if the model displayed racial and/or gender bias by replacing male pronouns with female pronouns and white with black and vice versa, and compared the probability distributions to those of the original sentences, as well as identified the 10 most frequent tokens per class, replaced them with synonyms, and observed if the predicted class or prediction probability distribution changed.

2 Related Work

2.1 GPT2-small

GPT-2, a transformer-based decoder-only model developed by OpenAI, builds on the success of the attention mechanism (Vaswani et al., 2017). The smallest version, GPT-2-small with 117M parameters, (fig.1) was initially designed as an unsupervised deep learning model to predict the next token given a document (Radford et al., 2019). But GPT-2 soon found its way into supervised machine learning tasks as well (Jain et al., 2024). GPT-2-small's architecture involves 12 scaled dot product attention layers, a layer norm, and a feed-forward network followed by another layer norm (the activations after the last layer norm are the final outputs of the transformer). It is trained with a massive corpus, with a vocabulary size of 50257. It has a model dimension (size of the embedding layer) of 768 and can calculate context for a maximum of 1024 tokens (Radford et al., 2019).

To perform classification with GPT, linear layer(s) are added after the output of the transformer block to take the transformer's prediction and map it to labels. In (Kheiri and Karimi, 2023), the authors combined three strategies of training GPT-3.5 Turbo for sentiment analysis (SemEval Task 4 dataset): prompt engineering, finetuning, and embedding classification. This study shows a 22% improvement in the F1 score compared to other emotion detection models.

2.2 Head Masking

Similar work focuses on using probing to identify "influential" attention heads in order to maximize performance, i.e. heads that have high accuracy and are relevant to a task, given that the amount of information captured by each head varies between layers, and more work remains on understanding what information is captured in each head (Jo and Myaeng, 2020). The head-masking portion of our analysis has a related objective, to find the attention heads most involved with emotion classification in GPT-2. Additionally, it has been shown that pruning attention heads in mBERT and XLM-R models generally improved performance on part-of-speech tagging, named entity recognition, and slot filling tasks, emphasizing how attention heads in transformers do not necessarily play equivalent roles in every task (Ma et al., 2021). While we are using a different model, we attempted to similarly investigate how much each head contributed to different emotion classification predictions, or if there was even a difference in contribution.

2.3 Token Analysis

Activation Patching (AP) is a technique used to analyze the internal structure of machine learning models. In (Heimersheim and Nanda, 2024), the authors explained methods and challenges of activation patching, which informed our design of the activation patching experiments. In (Meng et al., 2023), the authors introduce Rank-One Model Editing (ROME) to interpret the information stored in GPT-style models. By swapping corrupt activations with clean ones one at a time, the change in prediction highlights the importance of the patched component. While the original study shows which patched component is important across all layers in the model, we were more interested in looking at what changed between the two activations to cause the change in prediction, specifically at the attention layer.

In (Vig and Belinkov, 2019), the authors visualized the attention mechanism in GPT-2 to better understand where information was encoded - we attempted to similarly analyze the structure of GPT-2 by visualizing specific sentences, although in a more comparative fashion as we looked at activations before and after changing sentences to understand the impact of our changes on the model's predictions.

3 Data

We used the Twitter emotion detection dataset from Kaggle (kag). This dataset contains 393,822 total unique values with the following six classes and distributions - sadness (0) (121,187 values), joy (1) (141,067 values), love (2) (34,554 values), anger (3) (57,317 values), fear (4) (47,712 values), and surprise (5) (14,972 values). Sadness and joy are oversampled compared to the other classes, which will be discussed later in the report. Additionally, this is a preprocessed dataset, with punctuation and special characters removed.

4 Methods

4.1 Finetuning GPT-2

To finetune a GPT-2 for emotion classification, a pre-trained GPT2ForSquenceClassification model from HuggingFace was used. GPT2ForSquenceClassification is the GPT-2 transformer with a sequence classification head on top (linear layer). It utilizes the last token in order to do the classification. For finetuning, the original dataset was divided into a training and testing dataset with an 80/20 split. Training was done over 2 epochs, taking approximately 13 hours.

4.2 Interpretability Experiment 1 - Head Masking

After loading the baseline model $\mathcal{M}$ (section 4), a list of variant models $\mathcal{V}_i$ was obtained by iteratively applying a custom mask to the head "transformer.h.i.attn.c_proj", where i is the index of the head (i $\in$ [1, 12]), one at a time. We ended up with a list of 12 variant models, each with a different attention head masked, plus the initial baseline model.

$$\mathbf{A}_i = \mathbf{A}_i \times \mathbf{0} \tag{1}$$

Using PySpark for faster processing, 100 samples with the highest prediction confidence by the original model were extracted from each class and saved as a .csv file. These samples were run through the baseline and each of the variant models. For an input text x_j, if $\mathcal{M}$ predicts k_j, we were interested in finding the average difference in output probability between the variant model and baseline model, described by the following equation:

$$\text{Average}(P_{\mathcal{V}_i}(\hat{y_j}|\text{text} = k_j) - P_{\mathcal{M}}(\hat{y_j}|\text{text} = k_j))\forall j \tag{2}$$

The average difference in probability is indicative of how critical the masked head was to emotion

classification - high differences between the original and masked model meant the head performed a major role in classification, while smaller differences meant the head was less critical to the classification.

4.3 Interpretability Experiment 2 - Gender and Race Flipping

We wanted to investigate how different parts of GPT-2 contribute to the output through activation patching. Specifically, we wanted to investigate how gender and race impacts the output and where in the architecture encodes more information about gender and race.

To achieve this, every sentence that contained gender or racial terms was extracted and the terms were flipped. For example, female pronouns were replaced with male pronouns, "dad" was replaced with "mom", "white" was replaced with "black". We ran inference on the flipped samples and compared the probability distributions to the original predictions. We found the pairs of sentences (original and flipped) that caused the most significant difference in probability and observed their activations. The activation outputs of attention head 1 for each sentence were plotted as a heatmap and compared. Given our findings from head masking was that head 1 played the greatest role in prediction, we only observed the activations of attention head 1.

4.4 Interpretability Experiment 3 - Synonym Replacement

Finally, we found the 10 most common words of each class from the test dataset (excluding stopwords) using PySpark and replaced them with synonyms. For example, in the "fear" class, "scared" was replaced with "frightened". See table 3 in the appendix for the full list of replaced words in each class. After replacement, we ran inference on the dataset with the replaced words and compared probability distributions to the results from the original data.

5 Results

5.1 Finetuned GPT-2 Performance

After finetuning, GPT-2-small achieved a training accuracy of 0.945 and a test accuracy of 0.937 with a One-vs-Rest Train Area under the Curve (AUC) of 0.998 and test AUC of 0.992. Fig.2 shows the model's ability to predict the true class.

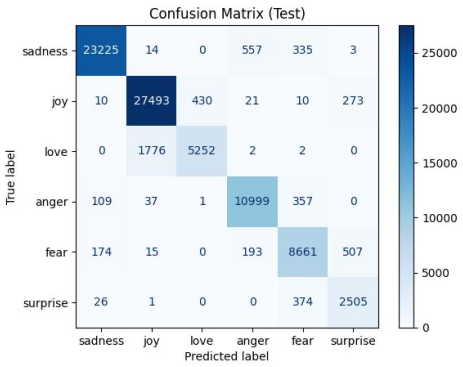

Figure 2: Confusion Matrix

5.2 Head Masking

The average difference in class probability between the original model and models with one attention head masked is shown in Figure 4. Masking head 1 changes the probability of the true class by 10% for "Surprise", 11% for "Anger", 12% for "Joy", 19% for "Fear", 28% for "Love", and -35% for "Sadness". Masking heads other than head 0 had minimal impact on the outputs, demonstrating that mostly head 1 was responsible for the emotion classification predictions. We found that masking all of the heads except head 1 also had a negligible impact on the model's predictions (Appendix Table 2). While the initial expectation was that each head would encode some property about the data, this experiment shows that for simpler tasks (like emotion classification), every head does not need to be involved.

5.3 Gender and Race Flipping

5.3.1 Gender Flipping

We found around 10 sentences in the test set that had at least 10% probability change in one of the classes after flipping the gender in the sentence. We chose 2 pairs of sentences that had the most significant probability difference before and after flipping gender, and the activations for those 2 pairs of sentences were plotted.

Fig.3 shows 6 heatmaps: each column is the heatmap for a pair of samples (original and gender-flipped). The first two rows correspond to activations of head 1 of the original sentence and the flipped sentence respectively. All 4 heatmaps are similar, with similar regions highlighted. This means that only a few tokens/dimensions in attention head 1 contribute to the encoding emotions. This further narrows down the number of tokens playing a role in classification from our conclusion

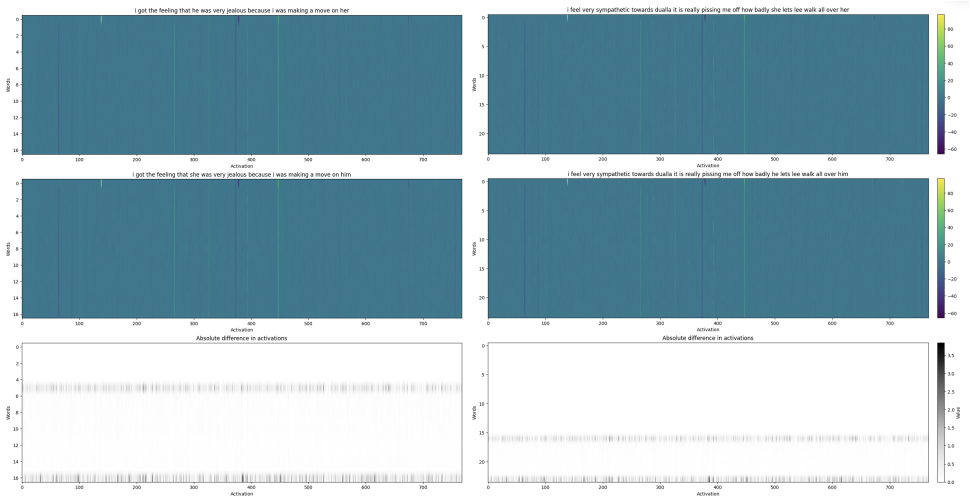

Figure 3: Heatmaps for both pairs of sentences (original and gender flipped). X-axis is the dimension of the attention head and y-axis is the words in the sentence.

in experiment 1. This is possibly due to the simple nature of the task, so only a handful of tokens are needed for accurate prediction.

The last row of heatmaps in fig.3, shows the absolute activation difference between the original and the gender-flipped sentence. Since the only difference between the two sentences is the gender, the activations indicate the encoding of gender information in the attention head. Between both pairs of heatmaps, there are common highlighted regions, like around dimension 380 and dimension 620. Tokens in these dimensions likely encode more gender information.

5.3.2 Race Flipping

After conducting the same methods for racial analysis, we did not find any sentences containing racial terms that led to significant probability changes after flipping the race. This indicates that GPT-2 has limited racial bias and/or the dataset did not contain many tweets that are racially biased.

5.4 Word Replacement

We found no significant differences in predictions between sentences with the original words versus sentences with the synonyms. This means GPT-2 was able to utilize the overall context of the sentences to conduct prediction and not overfit on common words.

This observation has an exception with sentences in class 5 (surprise) containing the original word “overwhelmed” (synonym "swamped") and the original word “surprised” (synonym "startled"). Sentences with these two words had significant probability changes after being replaced with their synonym. We hypothesize that this is because the synonyms chosen were not close enough, and the difference in meaning between the original word and synonym caused a shift in the meaning of the sentence, which led to the output being different.

6 Conclusions

In this paper, we dissected the GPT-2-small transformer to analyze its behavior after finetuning it on an emotion classification task. We show that only the first attention head drives the prediction given that emotion classification is a simpler task, and that GPT-2’s attention layers encode information about gender in some of its dimensions.

To expand on this project, we would further explore potential racial and gender bias, as that is a topic of interest for us and our current results are inconclusive. We might try to use a different emotion detection dataset and/or a different model and see if the results vary from our current findings.

Additionally, we would conduct the three experiments done in this paper (head masking, word flipping, and word replacement) using BERT and compare the results with GPT-2 to see the differences in architecture between the models. Due to our class imbalance, we would also retrain the model using a subset of each class to ensure the amount of data per class is equal.

7 GitHub

https://github.com/rjaditya-2702/CS6120-NLP

References

Twitter Emotion Dataset — kaggle.com. https://www.kaggle.com/datasets/adhamelkomy/twitter-emotion-dataset. [Accessed 07-12-2024].

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching.

Bhawna Jain, Gunika Goyal, and Mehak Sharma. 2024. Evaluating emotional detection classification capabilities of gpt-2 gpt-neo using textual data. In *2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 12–18.

Jo and Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. page 1956–1966.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

8 Appendix

Masked Head	Sadness	Joy	Love	Anger	Fear	Surprise
0	-0.351768572	0.127244908	0.281967689	0.115975701	0.191439452	0.101099012
1	-1.54972×10^{-8}	4.98308×10^{-9}	4.85383×10^{-9}	3.63388×10^{-9}	1.21184×10^{-9}	5.30456×10^{-11}
2	-1.43051×10^{-8}	1.85062×10^{-8}	7.05728×10^{-9}	-6.93551×10^{-9}	6.03665×10^{-9}	-5.04962×10^{-11}
3	1.31130×10^{-8}	1.11427×10^{-8}	-2.02861×10^{-9}	-3.77052×10^{-9}	4.10354×10^{-9}	2.27368×10^{-11}
4	3.69549×10^{-8}	2.91225×10^{-9}	6.08828×10^{-9}	-2.04219×10^{-9}	4.31132×10^{-9}	-1.58945×10^{-11}
5	-4.76837×10^{-9}	-8.40864×10^{-9}	4.12244×10^{-9}	1.53366×10^{-9}	-4.79900×10^{-9}	-6.03712×10^{-11}
6	-1.54972×10^{-8}	-9.60407×10^{-9}	-2.27120×10^{-9}	-1.74032×10^{-8}	-4.22294×10^{-9}	-1.43984×10^{-10}
7	3.57628×10^{-9}	-9.90606×10^{-9}	-9.30820×10^{-9}	-1.79252×10^{-8}	-1.33296×10^{-9}	-2.33730×10^{-10}
8	2.86102×10^{-8}	-1.06515×10^{-8}	-1.01126×10^{-8}	-1.78587×10^{-8}	-4.22118×10^{-10}	-1.39371×10^{-10}
9	1.78814×10^{-8}	-6.87358×10^{-9}	-8.26622×10^{-9}	-1.46728×10^{-8}	8.90771×10^{-10}	-1.36424×10^{-10}
10	2.02656×10^{-8}	-3.49728×10^{-9}	3.96860×10^{-9}	-1.24505×10^{-8}	3.99586×10^{-9}	9.82472×10^{-12}
11	6.43730×10^{-8}	-1.10244×10^{-8}	1.54352×10^{-8}	-1.73410×10^{-8}	-4.37200×10^{-9}	1.15809×10^{-10}

Note: Negative sign indicates, that the confidence $P_{\mathcal{V}_i}(\hat{y} = k|\text{text}) < P_{\mathcal{M}}(\hat{y} = k|\text{text})$

Table 1: Average difference of $P_{\mathcal{V}_i}(\hat{y} = k|\text{text})$, and $P_{\mathcal{M}}(\hat{y} = k|\text{text})$

Label	Difference
LABEL_0	$1.6808509833210473 \times 10^{-7}$
LABEL_1	$-1.0491646052496252 \times 10^{-8}$
LABEL_2	$1.7169386694604326 \times 10^{-8}$
LABEL_3	$-6.95903740677295 \times 10^{-8}$
LABEL_4	$1.0721862402363059 \times 10^{-8}$
LABEL_5	$-2.563566736668577 \times 10^{-10}$

Note: Negative sign indicates, that the confidence $P_{\mathcal{V}_i}(\hat{y} = k|\text{text}) < P_{\mathcal{M}}(\hat{y} = k|\text{text})$

Table 2: Average difference of probability of all head masked but head 0 variant w.r.t baseline's predicted class

Class	Original Word	Replaced Word	Class	Original Word	Replaced Word
Sadness	know	understand	Anger	know	understand
Sadness	time	period	Anger	people	individuals
Sadness	little	small	Anger	time	period
Sadness	people	individuals	Anger	little	small
Sadness	even	equal	Anger	want	desire
Sadness	still	nevertheless	Anger	angry	mad
Sadness	want	desire	Anger	think	believe
Sadness	think	believe	Anger	one	single
Sadness	bit	piece	Anger	even	equal
Sadness	life	existence	Anger	dont	avoid
Joy	time	period	Fear	know	understand
Joy	know	understand	Fear	still	nevertheless
Joy	people	individuals	Fear	little	small
Joy	one	single	Fear	bit	piece
Joy	want	desire	Fear	time	period
Joy	make	create	Fear	people	individuals
Joy	love	adore	Fear	think	believe
Joy	much	lots	Fear	scared	frightened
Joy	life	existence	Fear	want	desire
Joy	think	believe	Fear	afraid	fearful
Love	love	adore	Surprise	shocked	astonished
Love	know	understand	Surprise	strange	odd
Love	one	single	Surprise	overwhelmed	swamped
Love	people	individuals	Surprise	weird	uncanny
Love	time	period	Surprise	impressed	dazzled
Love	loved	adored	Surprise	amazed	astounded
Love	hot	boiling	Surprise	surprised	startled
Love	sweet	pleasant	Surprise	funny	hilarious
Love	little	small	Surprise	amazing	astounding
Love	loving	adoring	Surprise	curious	interesting

Table 3: The ten most common words per each class and their synonym replacements.

Original Text	Flipped Text	Probability Differences
"i got the feeling that he was very jealous because i was making a move on her"	"i got the feeling that she was very jealous because i was making a move on him"	[0.0097, 0.0955, -0.0138, -0.100, 0.0003, 0.0085]
"i feel very sympathetic towards dualla it is really pissing me off how badly she lets lee walk all over her"	"i feel very sympathetic towards dualla it is really pissing me off how badly he lets lee walk all over him"	[0.0052, 0.144, -0.075, -0.085, 0.00024, 0.0100]

Table 4: The original and flipped text and the differences in predicted probability for each class.