Prompt-tuned vs Fine-tuned models Which Better Account for Brain Language (and Vision) Representations?

Advaith Malladi
IIIT Hyderabad
2021114005

Patanjali Bhamidipati
IIIT Hyderabad
2021114014

Bhaiya Vaibhaw Kumar
IIIT Hyderabad
2019112021

***Abstract*—Previous research aimed to understand how the human brain processes language by examining brain responses to language input using pre-trained artificial neural network (ANN) models fine-tuned on Natural Language Understanding (NLU) tasks. However, full fine-tuning, which updates the entire parameter space, can distort pre-trained features, which is not consistent with the brain's robust multi-task learning ability. And in contrast, prompt-tuning freezes the pre-trained weights and only learns task-specific embeddings to fit a task. Could prompt-tuning produce representations that better match the brain's language processing than fine-tuning? Can we extend the task variety from NLU to Natural Language Generation tasks (NLG)? We explore these questions by comparing prompt-tuned and fine-tuned representations in neural decoding and encoding, which involves predicting linguistic stimuli from brain activities evoked by the stimuli and predicting brain representations from input stimuli. We notice interesting results in doing the above tasks for text and vision domain!**

***Index Terms*—prompt-tuning, fine-tuning, brain encoding, brain decoding**

I. Introduction

Previous studies have investigated brain encoding and decoding using fine-tuned models. However, full fine-tuning typically involves updating the entire parametric space, which can distort pre-trained features. This approach contradicts the way our brains robustly handle multi-tasking or learn new skills. In contrast, prompt-tuning freezes the weights of the pre-trained model and only learns task-specific embedding. This raises the question: could prompt-tuning produce representations that better capture the nuances of the brain's language processing compared to fine-tuning? If we end up with results either way, the outcome of this project would better brain alignment, with the most appropriate models. However, can the idea of prompt-tuning vs fine-tuning be extended to vision domain?

II. Fine-tuning

Fine-tuning refers to taking a pre-trained model and further training it on a new dataset, involving the entire model, including the initial layers. However, it is good to note that, since Language Models (LMs) already trained on huge data are being fine-tuned, the delta (δ) of the parameters' weight won't be that much.

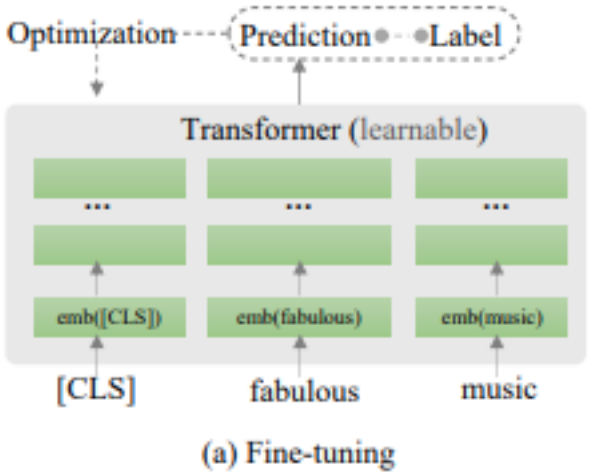

Fig. 1. Fine-tuning

III. Prompt-tuning

Prompt tuning [LARC21] is a technique in natural language processing (NLP) that adapts pre-trained language models to new tasks by training a small number of prompt parameters. It involves adding prompt matrix before the input text to guide the LLM towards generating the desired output. Only the *additional* prompt parameters are trained during the training part.

IV. Textual brain encoding and decoding

Given a text stimulus, obtaining the brain representations is encoding, while vice versa is decoding. Using the Pereira dataset, which includes sentences and their corresponding brain representations, we aim to generate final textual representations using both fine-tuned and prompt-tuned models across various tasks. The first question is which model to choose for fine-tuning and prompt-tuning. For this, we consider GPT-2 as our model, because it is a decoder-only model and can be extended to Natural Language Generation tasks easily. These models capture a wide range of linguistic knowledge and can be fine-tuned or prompt-tuned for specific tasks.

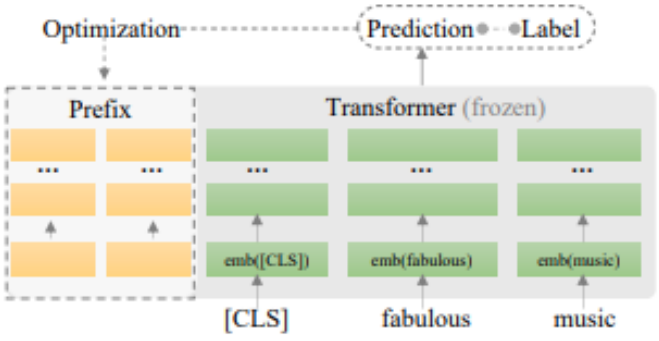

Fig. 2. Prompt-tuning

The second question is which tasks to select for fine-tuning or prompt-tuning. Since the original paper already worked on Natural Language Understanding tasks (NLU) using the BERT architecture, we tried to extend this by choosing diversified NLG tasks.
The tasks we chose are:

- Summarization
- Question-Answering
- Machine Translation

The specific reason for choosing these tasks are, when humans pick these skills, it takes diverse skill-set to actually learn each of these skill, thereby highlighting the novelty of the tasks. The other reason is, the original paper [SM23] which majorly had only classification datasets, so choosing these datasets and checking for the hypothesis would be a novel extension.

A. Methodology

We conduct both fine-tuning and prompt-tuning using GPT2 [YMG+23] on three different datasets: Machine Translation (MT), Summarization (Sum), and Question Answering (QA). Our objective is to train neural encoders and decoders to capture the correlation between brain representations and text representations. We are here training a Ridge Regression model for brain encoding and decoding which follows an L2 regularization. We perform this task on a region of interest (ROI) basis, focusing on specific brain regions, and then conduct an overall analysis. This procedure is carried out on two subjects, and we subsequently analyze the results across both subjects to validate our hypothesis. Throughout our analysis, we use the standard 2v2 accuracy metric to evaluate the performance of our models.

B. Results for Prompt-tuning and Fine-tuning

It is interesting to see how well the prompt tuned and the fine tuned models perform in the first place, only then there is a notion of brain encoding and decoding. So how do we evaluate? The approach we took is, we took every generated answer after prompt-tuning and fine-tuning and did a BERT-score [ZKW+20] with the ground truth of the dataset which is either the summary for summarization, answer for QA or translation for MT. The results looked like this:

TABLE I
BERT-SCORE FOR PROMPT-TUNING AND FINE-TUNING.

Task	Prompt-Tuned BERT Score	Fine-Tuned BERT Score
Summarization	0.67	0.64
Translation	0.58	0.60
Question Answering	0.69	0.69

A good BERT score indicates,a well worked out prompt-tuning and fine-tuning as above.

C. Metrics

The correlation between the decoded vectors and the ground-truth sentence embeddings is first computed. A successful matching is scored when the decoded semantic vectors have a higher degree of similarity with their corresponding brain activation patterns compared to alternative pairings. Formally, for each possible pair of sentence stimuli S_i and S_j, let X_{S_i} and X_{S_j} denote the brain images of neural responses to S_i and S_j. Let Z_{S_i} and Z_{S_j} denote the sentence embeddings produced by a tuned model, while D_{S_i} and D_{S_j} denote the decoded semantic vectors from brain images X_{S_i} and X_{S_j}. We **score 1** for a matching if

$$\text{corr}(D_{S_i}, Z_{S_i}) + \text{corr}(D_{S_j}, Z_{S_j}) > \text{corr}(D_{S_i}, Z_{S_j}) + \text{corr}(D_{S_j}, Z_{S_i}), \quad (1)$$

else 0.

D. Results

Consistently, Prompt-tuned GPT-2 better accounts for brain representations in both brain decoding and encoding than Fine-tuned GPT-2.Also, in Prompt-tuned models, the Summarization (Sum) and Question Answering (QA) models perform equally well, whereas the Machine Translation (MT) model perform less. Conversely, in Fine-tuned models, the MT model accounts better for brain representations than the other two tasks. Additionally, the results we obtained are consistent across subjects.
The analysis was performed across ROIs and it is worth noting that, each ROI gave results in compliance with our hypothesis showing that prompt-tuned models account better for brain representations than fine-tuned models.

TABLE II
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS (SUBJ1_LANG) IN THE LANGUAGE ROI FOR BRAIN DECODING

Task	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
Summarization	0.7548	0.6737
Translation	0.7094	0.6662
Question Answering	0.7548	0.6525

For compromising with the page limit and since, the original paper also focused on brain decoding, we are here attaching the decoding results of the brain. However, consistent trends were also followed in brain encoding as well.

TABLE III
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS (SUBJ1_TASK) IN THE TASK ROI FOR BRAIN DECODING

Task	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
Summarization	0.7406	0.6436
Translation	0.6971	0.6779
Question Answering	0.7406	0.6299

TABLE IV
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS (SUBJ1_VIS) IN THE VISUAL ROI FOR BRAIN DECODING

Task	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
Summarization	0.7530	0.6788
Translation	0.7090	0.6728
Question Answering	0.7530	0.6458

We can see the results from the language ROI above (theoretically important ROI) the prompt-tuned easily outperforming the fine-tuned models on all the three tasks. The other ROIs also show similar trends in results.

V. VISUAL BRAIN ENCODING AND DECODING

Given an image, obtaining the brain representations is referred to as encoding, while the reverse process is decoding. Utilizing the BOLD5000 dataset, which provides images and their corresponding brain representations, we aim to explore two key questions in the domain of vision. First, we need to determine which model to fine-tune and prompt-tune for vision tasks. Since the domain of vision is not at all explored, we chose the famous ViT-base (Vision transformer) [DBK+21] to study prompt-tuning vs fine-tuning on 2 datasets.

- MNIST - handwritten-number detection
- CIFAR-10 - General object detection, also involving scenic images, i.e. higher spatial complexity.

A. Methodology

We conduct both fine-tuning and prompt-tuning using ViT on the two different datasets: MNIST and CIFAR-10. Our goal

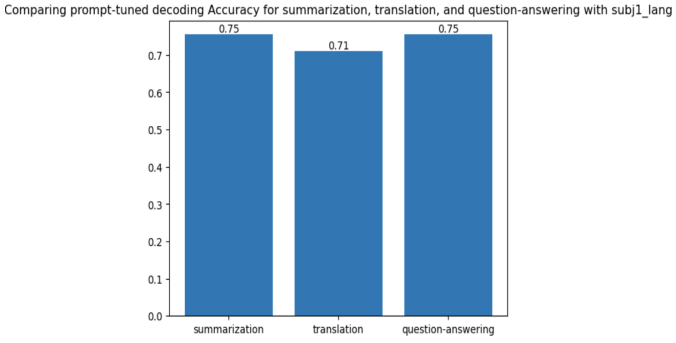

Fig. 3. Prompt-tuning for Language ROI across different tasks for Subject 1

TABLE V
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS (SUBJ1_DMN) IN THE DMN ROI FOR BRAIN DECODING

Task	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
Summarization	0.7181	0.6304
Translation	0.6618	0.6258
Question Answering	0.7181	0.6029

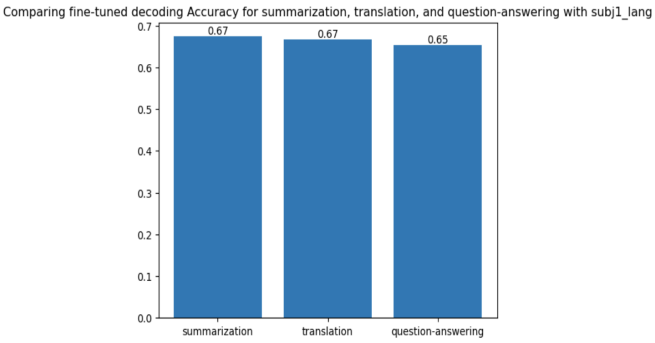

Fig. 4. Fine-tuning for Language ROI across different tasks for Subject 1

is to train neural encoders and decoders to capture the correlation between brain representations and text representations. We employ a Ridge Regression model for brain encoding and decoding, incorporating an L2 regularization. This task is conducted on a region of interest (ROI) basis, focusing on specific brain regions, followed by an overall analysis, this is all similar to the above methodology. But the images we get here are from BOLD-5000 dataset whose corresponding fMRI scans are also available in our dataset.

B. Prompt-tuning Vision models

In Language models, adding additional prompt-embedding matrix which represents the best embedding to solve the particular task makes sense. But how do we visualise this in the vision domain? It is exactly the same principle as how we do it in LMs. Even in LMs, while we are prompt-tuning we are really **not** cosidered with the what exactly the prompt represents, but we want the best representation of it to solve our task of prompt-tuning. Similarly, we are really not worried on what is exactly being represented in the prompt-embedding-matrix, but our end goal is that, this matrix should be the best representation for our task since these are the only trainable parameters. Additionaly we can look at this matrix as some embedding which triggers the vision classifier model to classify the image correctly. The word *trigger* is used lightly, but the idea is the same. Visual prompt-tuning from the paper [JTC+22] looks something like Figure 3:

C. Comparing Prompt-tuned vs Fine-tuned Vision models

As we have done above, we need to validate, the performance of our fine-tuning and prompt-tuning. Since the ViT-base is an encoder type model analogous to BERT, we can do a simple image recognition task on held-out test sets for

evaluation. On doing the above techniques, here is the table depicting accuracy(s).

TABLE VI
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED AND BASELINE VISION MODELS

Task	Accuracy
Base model	0.961
Prompt-tuning	0.972
Fine-tuning	0.975

Note that, all of them have undergone similar training procedure including the optimizer and a single epoch training. The above result is indicative of how any kind of tuning (prompt or fine tune) betters the base model for image classification. So that is why the study here is limited to Prompt-tuned vs Fine-tuned.

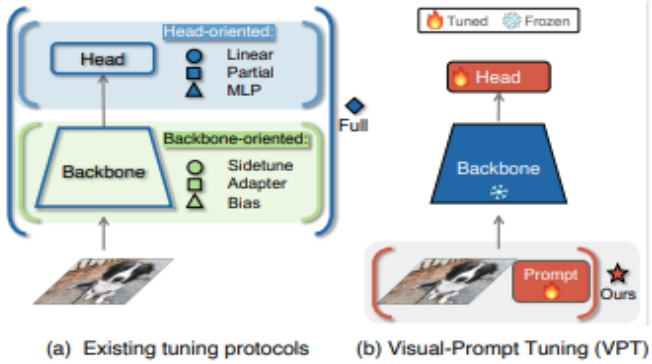

Fig. 5. Visual Prompt Tuning

D. Results

To our surprise, we found that fine-tuning models consistently outperform prompt-tuning models. The reason behind this phenomenon remains unknown, although prompt-tuning models still exhibit a similar trend to the text data. It was also evident from the results that ViT fine-tuned on CIFAR-10 performed better than ViT fine-tuned on MNIST for both *brain decoding* and *brain ecoding*. We can hypothesize that, maybe because of the diverse nature of the dataset, it has images which have more complex spatial dimensions and color dimensions, hence can account for better brain representations rather than a much simpler dataset of gray-scaled numbers.

Dataset	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
MNIST	0.756	0.824
CIFAR	0.754	0.819

TABLE VII
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS USING LOC

VI. CONCLUSION

Which model accounts for brain representations better? This question has been floating around the cognitive research for long. Once the studies started with Artificial Neural Networks (ANNs), it went on to explore Neural Taskonomy [OAA+22] and then to use several fine-tuned models for this task. Next came in the timeline [SM23] which explored how BERT [DCLT19] fine-tuned and prompt-tuned on 10 NLU (here: classification) tasks showed its prompt-tuned version to be a better brain representing version. We, via this project extend this idea into 2 unexplored paths.

One, by delving into Natural Language Generation tasks, replacing the existing NLU tasks. Two, by delving into the vision domain and checking for the same idea.

Results show that, even when extending for generation tasks in text, the results still align with the original paper, which is also indicating a task-independent edge of prompt-tuning over fine-tuning in text based models for accounting for brain representations.

When the similar idea was applied to the Vision domain, the results flipped! Indicating a better brain representations by fine-tuned vision models over prompt-tuned. Though the reason is not immediately decipherable by existing theoretical frameworks, we leave this result as an experimental result and to be extended to various other vision tasks, as this stands the starting point of the research henceforth.

Dataset	Prompt-Tuned Accuracy	Fine-Tuned Accuracy
MNIST	0.750	0.836
CIFAR	0.748	0.821

TABLE VIII
COMPARISON OF ACCURACY FOR PROMPT-TUNED AND FINE-TUNED MODELS USING OPA

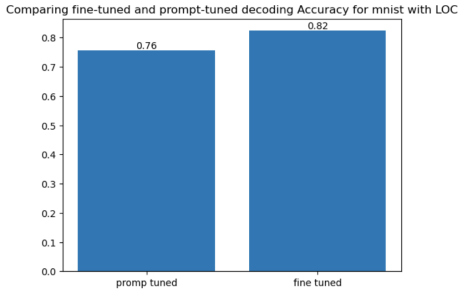

Fig. 6. Prompt-tuning vs Fine-Tuning for LOC ROI

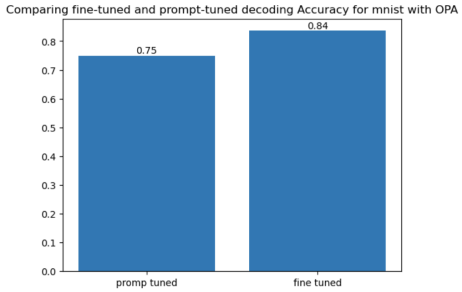

Fig. 7. Prompt-tuning vs Fine-Tuning for OPA ROI

REFERENCES

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[JTC+22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022.

[LARC21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[OAA+22] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States, July 2022. Association for Computational Linguistics.

[SM23] Jingyuan Sun and Marie-Francine Moens. Fine-tuned vs. prompt-tuned supervised representations: Which better account for brain language representations?, 2023.

[YMG+23] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, 2023.

[ZKW+20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.