Linear Transformation Captures Change in LLM Representation Space during Finetuning, but not Mapping across Model Types

Shan Gao
University of Chicago

Lihao Sun
University of Chicago

Jialing Jiao
University of Chicago

Abstract

In this study, we investigate the potential of affine transformations to bridge the representational gap between weak and strong models in the context of weak-to-strong generalization. Previous research has demonstrated that strong models fine-tuned on noisy labels can surpass weak models fine-tuned on ground truth labels, yet still exhibit a performance gap relative to strong models fine-tuned on ground truth labels. Our work explores whether it is possible to utilize the ground truth task representations from a weak model to enhance the performance of a strong model through affine transformations. We examined two approaches: between-model and within-model transformations. For the between-model transformation, we attempt to map the ground truth task representation from the weak model's representation space directly to that of the strong model using an affine transformation matrix. The within-model transformation focuses on extracting task representations changes from the weak model's training dynamics, analyzing the changes pre- and post-fine-tuning. Our preliminary results reveal that affine transformations can capture within-model conceptual shifts to match the finetuning level, they struggle with between-model transformation to fully capture the complex mappings required between these models' representation spaces. This suggests the need for alternative methodologies that might better accommodate these discrepancies.

1 Introduction

Weak-to-strong generalization [2] highlights the phenomenon that computationally *strong models* fine-tuned on *noisy, weak labels* surpass the performance of computationally *weak models* fine-tuned on *ground truth labels*, but still possess a performance gap (PG) compared to *strong models* fine-tuned on *ground truth labels*. Probing studies by [2] have shown that, even though the strong models are fine-tuned on noisy labels, the salience of ground truth task representation has in fact increased during the fine-tuning process, and the probing performance exceeded the performance observed directly from model behaviors. These findings indicate that strong models may be able to "infer" the ground truth task representation from noisy training samples to some extent, but not use this representation to its full potential when generating outputs.

This naturally leads to the following question: Can we find the ground truth task representation in the strong model's representation space using some reference points, and use it directly for generating model outputs to increase the performance gap recovered (PGR)?

In this project, we considered two complementary approaches in a preliminary exploration of the question: (1) between-model transformation: mapping the ground truth task representation in weak

Preprint. Under review.

model's representation space (given that the weak model has access to ground truth labels by the problem setup) to the strong model's representation space; (2) within-model transformation: extracting emergent task representation from the training dynamics via pre- and post-fine-tuning comparison within one model (referring to the strong model in the original weak-to-strong setup, but this approach itself is not model-specific).

Here, as a first step, we empirically tested the plausibility of modeling both between-model transformation and within-model transformation as *linear* transformations. We studied the last layer activation (h.11) in open-source GPT2 model series on three datasets, including a sentiment classification task with the amazon polarity dataset, a scientific question answering dataset SciQ, and a reward modeling dataset HH-RLHF. [4, 1]

2 Representation space between-model transformation

In the weak-to-strong generalization problem setting, the weak model has access to ground truth labels whereas the strong model does not. In this approach, we tested the possibility of transforming the ground truth task representation from the weak model's representation space to the strong model's representation space via an affine transformation.

2.1 Method

We formalized the problem of finding a linear transformation between two models' representation spaces as follows:

$$argmin_A \mathbb{E}(\|A\lambda^{(weak)}(x) - \lambda^{(strong)}(x)\|)$$

where A denotes the affine transformation matrix from the weak model's representation space to the strong model's representation space we'd like to find; $\lambda^{(weak)}(x)$ denotes the weak model's last layer representation of each data point; $\lambda^{(strong)}(x)$ denotes the strong model's last layer representation of the corresponding data point. We used GPT2 as the weak model, GPT2-XL as the strong model, and 10k training data points from the amazon polarity dataset to find the A that minimized this loss function via gradient descent.

After obtaining the affine transformation matrix A, we applied it to the weak model's ground truth task representation (found via probing) to infer the ground truth task representation in the strong model:

$$\hat{\mathbf{w}}^{(strong)} = A\mathbf{w}^{(weak)}$$

We tested whether $\hat{\mathbf{w}}^{(strong)}$ can be used to classify samples based on their representations in the strong model by computing their signed projection onto the $\hat{\mathbf{w}}^{(strong)}$ vector:

$$pred\_label = sign(\lambda^{(strong)}(x) \cdot \hat{\mathbf{w}}^{(strong)})$$

We compared $pred\_label$ to the ground truth label to determine the sentiment classification accuracy.

2.2 Result

We consider the representation spaces of GPT2 ($\mathbf{w}^{(weak)}$) and GPT2-XL ($\mathbf{w}^{(strong)}$) on the Amazon Polarity dataset. In Figure 1 we visualize cross-model representation spaces using t-SNE, indicating potentially linear relationships.

We found that the linear probe accuracy of $pred\_label$ against ground truth labels was 0.5025, indicating that linear transformation does not capture mapping across GPT2 and GPT2-XL models' representation spaces despite the similarity in visualization.

3 Representation space within-model transformation

The second approach is to look at the task representation extraction problem from the perspective of vertical comparison across time, i.e., pre- and post-fine-tuning between which the target task representation emerges / becomes more salient.

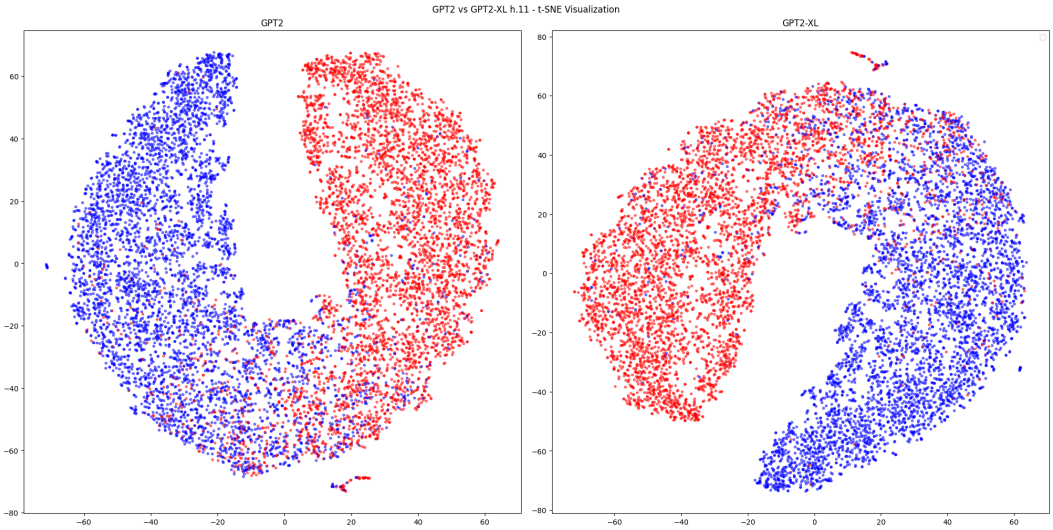


Figure 1: Cross-model comparison between GPT2 and GPT2-XL showing correspondence between representation spaces

For the scope of this project, we would like to first ascertain whether the pre- and post-fine-tuning representation space change can be modelled by a *linear* transformation. Therefore, we utilized ground truth labels when constructing our loss function, even though in the original weak-to-strong setup, ground truth labels are presumably inaccessible for the model from which we would like to extract the task representation from. Some preliminary experiments with weak labels were also conducted, but we defer to future studies to look for ways of constructing loss functions without introducing ground truth labels.

3.1 Method

We formalized the problem of extracting emergent task representation from the training dynamics as follows:

$$argmin_{A,\delta}\mathbb{E}(\|(A\lambda(x) + label \times \delta) - \tilde{\lambda}(x)\|)$$

where δ is the rank-one task representation learned through the process of fine-tuning that we are interested in extracting; A is an affine transformation matrix; $\lambda(x)$ denotes the last layer model representation of each data point pre-fine-tuning; $\tilde{\lambda}(x)$ denotes the last layer model representation of each corresponding data point post-fine-tuning.

If we are able to find an informative δ, it will indicate that the geometry of representation spaces of the same model pre- and post-fine-tuning are off only by an affine transformation, and that different classes become more linearly separated through fine-tuning by shifting to or away from the δ direction, which is, by definition, the task representation we are looking for. We tested this hypothesis on both GPT2 and GPT2-XL models respectively, again using 10k training data points from the amazon polarity dataset to find the A and δ that minimized the loss function via gradient descent.

We tested whether δ can be used to classify samples based on their representations in the post-fine-tuning model $\tilde{\lambda}(x)$ by computing their signed projection onto the δ vector:

$$pred\_label = sign(\tilde{\lambda}(x) \cdot \delta)$$

We compared $pred\_label$ to the ground truth label to determine the sentiment classification accuracy using δ.

We constructed the representation space $\hat{\lambda}$ from the base GPT-2 representations λ using A and δ. To measure the task saliency boost of the resulting A and δ, we trained a linear probe and tested it against the labels:

$$\hat{\lambda}(x) = A\lambda(x) + label \cdot \delta$$

This pipeline is illustrated in 2.

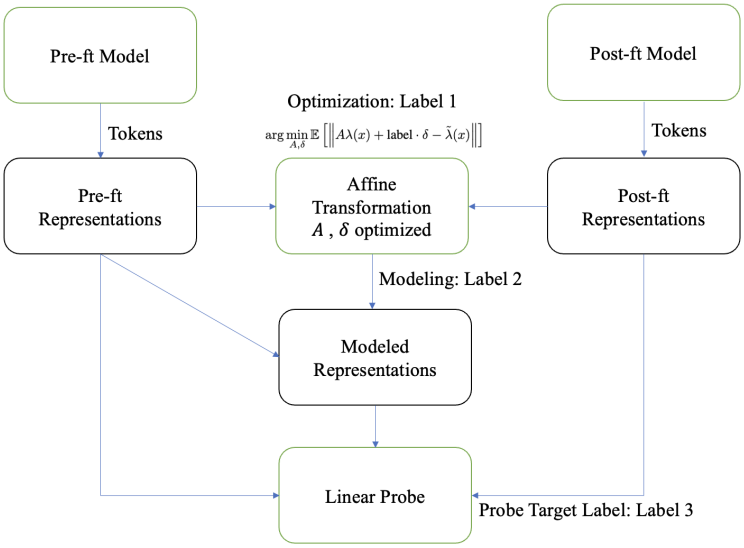


Figure 2: Within-model Transformation Pipeline.

3.2 Result

Our findings suggest that the sentiment information is already significantly encoded in the GPT2-XL model, as evidenced by the lower classification accuracy after the transformation, with an accuracy of only 0.4904. In contrast, the GPT2 model, which showed a projection classification accuracy of 0.8994, indicates that there is still substantial potential for improvement through fine-tuning. This is further supported by the comparison to the GPT2's projection classification using the probe vector direction, which resulted in a slightly lower accuracy of 0.8790.

3.2.1 Visual Analysis

To gain more intuition for changes in the representation space during fine-tuning, we employed PCA and t-SNE to visualize the activations.

In particular we observed that separation becomes stronger in the later layers of the transformer. To that end, we chose to focus on layer $h.11$ in the ensuing experiments.

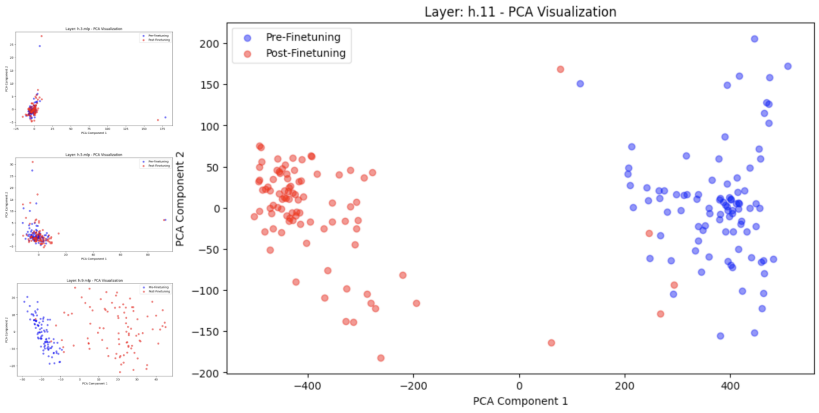
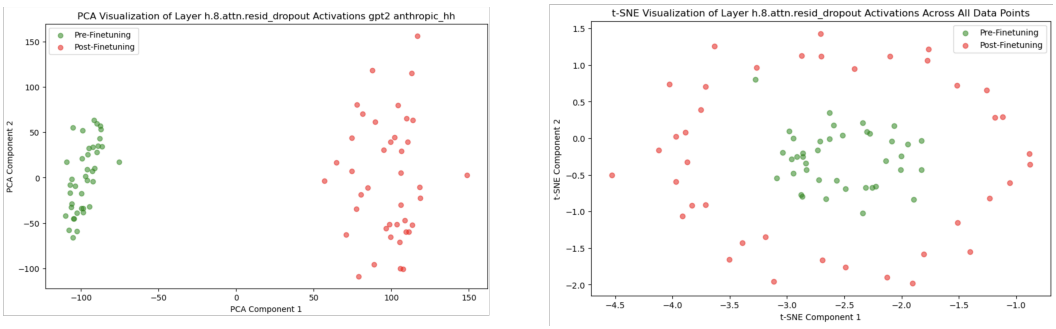


Figure 3: PCA analysis on activation change pre- and post- fine-tuning for one data point in the Amazon Polarity dataset. On the left are layers 3,5, 7 respectively and on the left is layer 11.



(a) PCA (b) t-SNE

Figure 4: PCA and t-SNE analysis on activation before and after finetuning with Anthropic HH dataset. Each dot here represents a datapoint.

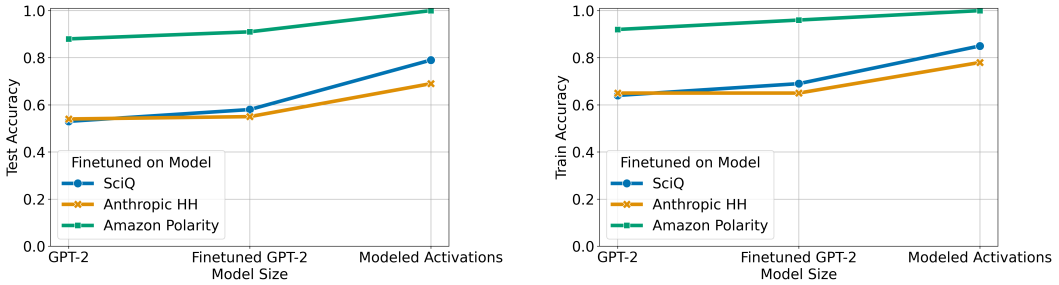
3.2.2 Modelled Activation using Groundtruth Labels $\hat{\lambda}_{gt}$

As a proof of concept, we optimized for the affine transformation matrix A and delta shift δ on signed ground truth labels for the SciQ, Amazon Polarity, and Anthropic HH datasets (with Labels 1, 2, 3 being ground truth labels in 2). We constructed $\hat{\lambda}_{gt}$ based on signed ground truth labels and trained a linear probe to test the performance of $\hat{\lambda}_{gt}$ against the ground truth labels.

Dataset	Model	Train Accuracy	Test Accuracy
SciQ	Base GPT-2	0.64	0.53
	Finetuned GPT-2	0.69	0.58
	Modeled Activations $\hat{\lambda}_{gt}$	0.85	0.79
Amazon Polarity	Base GPT-2	0.92	0.88
	Finetuned GPT-2	0.96	0.91
	Modeled Activations $\hat{\lambda}_{gt}$	1.00	1.00
Anthropic HH	Base GPT-2	0.65	0.54
	Finetuned GPT-2	0.65	0.55
	Modeled Activations $\hat{\lambda}_{gt}$	0.78	0.69

Table 1: Train and test linear probe accuracy for different datasets.

As we had anticipated, the task saliency, measured by linear probe accuracy, improved by more than 10 percentage points across all datasets. This is because $\hat{\lambda}_{gt}$ was construct to improve class separation along the labels.



(a) Test Accuracy (b) Training Accuracy

Figure 5: Linear Probe accuracy on three datasets: SciQ, Anthropic Helpfulness, and Amazon Polarity.

3.2.3 Modelled Activation using Weak Labels $\hat{\lambda}_{weak}$

To test our hypothesis, we focused on weak labels generated by GPT-2 fine-tuned by the Anthropic HH reward modeling dataset (with Label 1 and 2 being weak labels and 3 being the ground truth label in 2). We optimized for A and δ on the weak labels and constructed $\hat{\lambda}_{weak}$ based on the inferred representation. We used a linear probe to measure the performance of $\hat{\lambda}_{weak}$ against ground truth labels.

Our observations indicate that linear probe accuracies are retained or even improved when compared to the performance of fully finetuned models. This suggests that affine transformations are capable of capturing shifts in learned representations effectively, achieving outcomes comparable to those of finetuning. However, in order to apply this to a stronger model with higher dimensions, we still need robust mapping strategy as described in 2.

Dataset	Model	Train Accuracy	Test Accuracy
Anthropic HH	Base GPT-2	0.66	0.53
	Finetuned GPT-2	0.68	0.55
	Modeled Activations $\hat{\lambda}_{weak}$	0.69	0.55

Table 2: Linear probe train and test accuracies for different datasets. Weak labels are used for modeled Activations.

4 Discussion

Our work can be framed as a subproblem in interpretability, specifically the linearity hypothesis. Past studies have found that concepts such as gender, deception, and toxicity are linearly embedded in the representation space of a transformer [3, 5]. In this paper, we sought to extend this hypothesis to explore cross-model linearity and the linear representation of tasks.

Our preliminary findings indicate that within-model affine transformations can capture certain learned concepts. However, transitioning these transformations from models with lower-dimensional representation (e.g., GPT-2) to those with higher dimensions (e.g., GPT-2 XL) presents challenges due to the inability of linear transformations to effectively map across their representational spaces. This suggests a need for the development of alternative methodologies that can accommodate the discrepancies in dimensionalities between different models.

For transferring within-model transformations from weak to strong models, we would also like to experiment with the following approach: Decompose the pre-finetuning representations (X) of the strong model into principal components. Project X onto the subspace defined by the top n eigenvectors to produce a new matrix Y in n-dimensional space, where $Y = XP$ and P is the matrix composed of the selected eigenvectors. Subsequently, apply the within-model affine transformation obtained from the weak model, resulting in $Y' = AY + \text{weak\_label} \times \delta$. Finally, transform back into the original strong model representation space using $X' = Y'P^T$. Then test linear probe on X' and compare with the accuracy given by strong models finetuned on weak labels directly.

This method leverages dimensionality reduction and projection techniques to potentially transfer the ground truth finetuning effects in the weak model to the strong model.

Furthermore, our study was limited to three specific tasks. Future research should expand upon this by incorporating a wider array of datasets to further validate and refine the findings.

Lastly, while affine transformations have served as the primary methodological approach in our experiments, other techniques, such as neural network architectures, PCA reconstructions, and non-linear mapping methods, should also be explored. These approaches may offer more robust ways of modeling the transformations needed to bridge the dimensional gaps between various model architectures. For example, for direct between-model transformations, one possibility could involve training a neural network to map between the pre-finetuning representations of weak and strong models. Subsequently, this neural network could be utilized to transform the post-finetuning representations from the weak model to generalize to representations of the strong model.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Asakell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[3] A. Li et al. Gender representation in transformer models. *Journal of Artificial Intelligence Research*, 1(1):1–14, 2024.

[4] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

[5] B. Zhao et al. Toxicity and deception in transformer models. In *Proceedings of the Neural Information Processing Systems*, volume 36, pages 1–12, 2023.