Effects of Pre-training and Fine-tuning Time on the Linear Connectivity of Language Models for Natural Language Inference

Stanford CS224N Custom Project

Kushal Thaman
Department of Computer Science
Stanford University
kushalt@stanford.edu

Abstract

It is widely known that two neural networks, when trained similarly on the same data under different random seeds converge to a similar loss when linearly interpolated between each other in parameter space, a property known as *linear mode connectivity* (LMC). We study whether Large Language Models for text classification (on datasets such as QQP, MNLI, CoLA) optimize to the same, linearly connected minima under different samples of training time (e.g. pre-training time or fine-tuning steps). Understanding the level of sensitivity of previous LMC analysis to the amount of pre-training or fine-tuning these networks have undergone would reveal interesting implications about our ability to predict losses of interpolated models and out-of-distribution generalization by appropriately controlling the number of training or fine-tuning steps. In our experiments, we find that (a) for a pre-trained RoBERTa model, more fine-tuning leads to less convexity gap (CG) between different random seeds of the model (higher LMC), though the convexity gap is still large (much larger than 0, which would imply total LMC), and (b) for a given amount of fine-tuning on text classification tasks, highly pretrained RoBERTa models display more LMC, or less path dependence on the training processes than the smaller, base RoBERTa model. In this way, we show our hypotheses that (a) pre-trained LLMs produce higher *linearly connectivity* as they are fine-tuned on a task complementary to the pre-training task, and that (b) fine-tuning overtrained models produces more LMC seeds, though this is not necessarily true for models that are not trained much past convergence.

1 Key Information to include

- Mentor: Nelson Liu, Prof. Tatsu Hashimoto

2 Introduction

When neural networks are trained with Adam or mini-batch Stochastic Gradient Descent (SGD), with the same training setup, hyperparameters, and initialization, the training examples are presented to the learner in a random ordering at each step of an epoch. Further, the random initialization can also cause certain features in the network to be encoded in the network in different and unexpected ways (Lecomte et al., 2023). This randomness causes the trained network to behave slightly differently from training run to training run, altering its trajectory through the loss landscape until it reaches the minima. Thus, different random seeds of the same model could reside at different points (though usually close) in the loss landscape. Whether these points lie in an entirely different loss basin is an interesting question to consider.

Prior work has demonstrated that fully trained neural networks (both in computer vision and natural language) often have a tendency to fall into a single linearly-connected loss basin across different training runs (with the exact same setup except using different random seeds), implying that it is possible to traverse a linear path in the parameter space such that the model loss stays non-incremental (roughly constant or decreasing) along that path (Garipov et al., 2018). Some of these works have presented this property, called linear mode connectivity (LMC) in popular literature, as a fundamental inductive bias of SGD linked to in-domain (ID) generalization. While demonstrating these results for image and text classification has shown results regarding the trained network's inability to access multiple basins, prior work has only considered such analysis over a fixed, trained model, as opposed to several training checkpoints of the model. Besides pre-training time, it is also interesting to consider whether a pre-trained model's random seeds become more or less similar as it is fine-tuned on a task complementary to the pre-training task.

In this work, we study LMC in text classification tasks on BERT and the RoBERTA-base and RoBERTa-large models Liu et al. (2019). We focus on Natural Language Inference (NLI) via text classification, a similar task as chosen by Juneja et al. (2022). We choose this in part because text classification is a widely applicable task that pre-trained language models are commonly fine-tuned on, and partly because Juneja et al. (2022) demonstrates a result in the paper that RoBERTa models achieved much higher LMC than BERT models in their experiments (see Appendix E of Juneja et al. (2022)). Our contributions are as stated below:

- We **propose a novel model similarity metric to measure the strength of LMC between random seeds**. Built on previous work . Across all our experiments, we find consistency between CG and ICG, validating the theoretical foundations of the metric.
- For fixed, pre-trained BERT ('bert-base-uncased') and RoBERTa models ('roberta-base' & 'roberta-large'), more fine-tuning on the Quora Question Pairs (QQP) Dataset (an NLI task) and the Multi-Genre Natural Language Inference (MNLI) Dataset leads to higher linear mode connectivity between random seeds. We measure LMC via a previously defined metric of model similarity, called convexity gap (CG), as well as a novel model similarity metric to measure LMC proposed by us, called the 'Improved Convexity Gap' (ICG). For both BERT and RoBERTa, we obtain decreasing values of CG and ICG, indicating **higher LMC between** 2 **random seeds with increasing fine-tuning steps**.
- The overall CG and ICG values for BERT are still fairly large (much larger than 0, which would imply total LMC). The CG and ICG values for RoBERTa-base are smaller than corresponding BERT values, and the values for RoBERTa-large are even smaller. This indicates that for a given amount of fine-tuning time on NLI tasks (QQP & MNLI), **models that have undergone higher pre-training will emerge as more LMC** than models that have not been trained past convergence.

3 Related Work

The earliest work that showed that models discovered using SGD or optimization algorithms built on top of SGD are connected by the linear paths over which loss stays the same or decreases was presented by Draxler et al. (2018). Most of the past literature exclusively studied mode connectivity in models trained on image classification tasks, and the very little literature studying models outside of image classification failed to find meaningful results in those methods which relied on LMC (Garipov et al., 2018). Juneja et al. (2022) showed that a pretrained model may not commit exclusively to a single basin, but instead favor a small set of them, challenging the commonly-held view on LMC in image classification. Juneja et al. (2022) challenges the general claim that SGD consistently finds the same linearly-connected loss basin in every training run. It aims to understand how the large barriers of increasing loss in the linear path along the weight can relate to different generalization strategies in text classification. This line of work also has implications on ways to build models that demonstrate better generalization capabilities, e.g. by model souping (Wortsman et al., 2022). Wortsman et al. (2022) pursued a similar idea as Juneja et al. (2022) in finetuning a natural language model with a linear classifier head, developiong weight ensembles (called model soups) that depend on assumptions of LMC. They found that model souping was much less effective in the text classification setting compared to the image classification one.

4 Approach

4.1 Model Setup

We use the pre-trained BERT ('bert-base-uncased') and RoBERTa models ('roberta-base' & 'roberta-large') with a sentence classification head (as the intended task is paraphrasing (QQP) or semantic-validity NLI (MNLI)). As we were logging our LMC metrics every 1000 fine-tuning steps, we did not use 'FacebookAI/bert-base-uncased-mnli', 'FacebookAI/roberta-base-mnli' or 'FacebookAI/roberta-large-mnli', and instead fine-tuned our own BERT and RoBERTa models (no checkpoints were available).

The reason we picked BERT and RoBERTa analysis is that it provides an excellent suite of models with an increasing amount of pre-training time. RoBERTa is trained with much larger mini-batches and learning rates than BERT (108M parameters), as well as being trained on $10\times$ more data as well as trained past convergence. RoBERTa-base is 123M parameters as opposed to 354M parameters for RoBERTa-large.

4.2 Linear Mode Connectivity

We define linear mode connectivity as the property of a network $f(., \theta)$ such that for two random seeds with weights θ_1 and θ_2 (the only difference in training being data ordering or random seeding), there exists an $\alpha \in [0, 1]$ such that the following inequality for the loss function $\mathcal{L}(f(., \theta)) = L(\theta)$ is satisfied:

$$L(\alpha\theta_1 + (1 - \alpha)\theta_2) \leq \alpha L(\theta_1) + (1 - \alpha)L(\theta_2)$$

As in Juneja et al. (2022), we can also study which 'loss basin' the model sits in, a measure that is predictive of the model's generalization behavior. To do this, Juneja et al. (2022) improves upon the Barrier Height metric, that aims to find the line of maximum possible loss difference between interpolated models and the weighted average of the two random seed models:

$$BH(\theta_1, \theta_2) = \sup_{\alpha} [L(\alpha\theta_1 + (1 - \alpha)\theta_2) - (\alpha L(\theta_1) + (1 - \alpha)L(\theta_2))] \quad \alpha \in [0, 1]$$

Juneja et al. (2022) does this by considering sub-segments of each such maximal-loss-difference line:

$$CG(\theta_1, \theta_2) = \sup_{\gamma,\beta} BH(\gamma\theta_1 + (1 - \gamma)\theta_2, \beta\theta_1 + (1 - \beta)\theta_2) \quad \gamma, \beta \in [0, 1]$$

An intuition for why pre-training or supervised fine-tuning time should affect these properties is that the model is pushed further deeper into the basin as it learns to generalize better by undergoing more pre-training. From the defintion above, as $CG(\theta_1, \theta_2) \to 0$, the supremum of the difference between the loss on the interpolated linear path and the weighted average of the loss of the two random seeds converges to 0, implying that the two random seeds are completely path independent (of the training processes) and reside deeply in a linearly connected basin. The same logic follows for $BH(\theta_1, \theta_2)$. Below, we propose a novel metric to better extend the convexity gap metric; while we are yet to empirically test this improved metric, this novel contribution should help better capture semantic generalization behavior by calculating the best sub-variations of convexity for each sub-segment.

$$ICG(\theta_1, \theta_2) = \sup_{\sigma,\delta} CG(\sigma\theta_1 + (1 - \sigma)\theta_2, \delta\theta_1 + (1 - \delta)\theta_2) \quad \sigma, \delta \in [0, 1]$$

In our two-stage experimental setup, we proceed to test LMC as follows:

- For the BERT and RoBERTa models, we fine-tune each of the 3 models (with two random seeds, namely 'seed(42)' and 'seed(43)') on the Quora Question Pairs (QQP) Dataset (an NLI task that tests paraphrasing capability), logging a checkpoint (to test LMC) at every $20,000$ training steps. leads to higher linear mode connectivity between random seeds.
- At the end of the fine-tuning, we use the logged Barrier Height (Juneja et al., 2022), CG and ICG values for BERT and RoBERTa models. We compare them to the ideal case of total LMC, which would give $BH, CG, ICG \to 0$.

5 Experiments

5.1 Data

We used a fixed but uniformly randomly chosen 10% of the Quora Question Pairs (QQP) Dataset and the Multi-Genre Natural Language Inference (MNLI) for fine-tuning the BERT and RoBERTa models. QQP contains over 400,000 question pairs, and each question pair is annotated with a binary value indicating whether the two questions are paraphrase of each other. MNLI is a collection of 433k sentence pairs annotated with textual entailment information. We chose a subset of the dataset to save money on compute, as doing every random seed for each of our 3 models in consideration would cost significantly more. Past literature in connectivity has focused on classification tasks in text and vision, making binary relationship datasets like QQP and MNLI natural choices for our setting. We decided on a natural $80:20$ split for training and test dataset for both MNLI and QQP.

5.2 Evaluation method

Since we only fine-tuned our models on a small subset of MNLI and QQP, it is difficult to use F1 scores and accuracies for those as a baseline to assess the quality of our fine-tuned models. Thus, for evaluation, we refer to Juneja et al. (2022) and use the $CG \Rightarrow 0$, $ICG \Rightarrow 0$ and $BH \Rightarrow 0$ as a baseline to confirm the hypotheses that more pretraining leads to stronger mode connectivity and more convex basins. Further, given the monotonic nature of the three functions, we can reasonably assume that lower values of these metrics imply a stronger independence from details of the training process, as well as a stronger connectivity of basins in the linear path interpolation. We also define our own evaluation metrics, namely Δ BH, Δ CG, and Δ ICG, arguing that more (higher) negative values of these will imply a stronger notion of LMC in the random seeds of a text classification model. As our baselines, we will use the LMC metric values for fully-trained but non-fine-tuned BERT and RoBERTa models.

5.3 Experimental details

For consistency, used the following hyperparameters for both BERT ('bert-base-uncased') and RoBERTa ('roberta-base' & 'roberta-large) fine-tuning on QQP and MNLI:

- Dataset batch size: 32
- Optimizer: Adam (weight decay = 0.1, betas = $(0.90, 0.98)$, $= 1 \times 10^{-6}$)
- Learning rate: 1×10^{-5}, no clipping of norms, polynomial decay with 7432 warmup updates
- Fine-tuning time: $1, 60, 000$ training steps
- Dropout: 0.1, attention dropout = 0.1

We load the pre-trained weights for each of these from HuggingFace. For compute needed for fine-tuning, we used a A100 GPU.

5.4 Results

In Figure 1, we pick one random seed of the RoBERTa-base model to visually demonstrate the fine-tuning loss, validation loss, accuracy on MNLI test set, and F1 score for the QQP test set over the course of fine-tuning.

We report the ΔLMC metrics for each of our models fine-tuned on QQP in Table 1 and those fine-tuned on MNLI in Table 2.

- First, we note that the ΔICG we initially proposed appears to have consistent results as the other metrics, supporting its validity in performing a similar underlying evaluation task.
- We found that $\Delta(BH)$ had the biggest jump between start and end of fine-tuning. This is expected, as the barrier height does not capture the individual segments or sub-segments of the maximum loss difference between interpolated model and the weighted average of the two models.

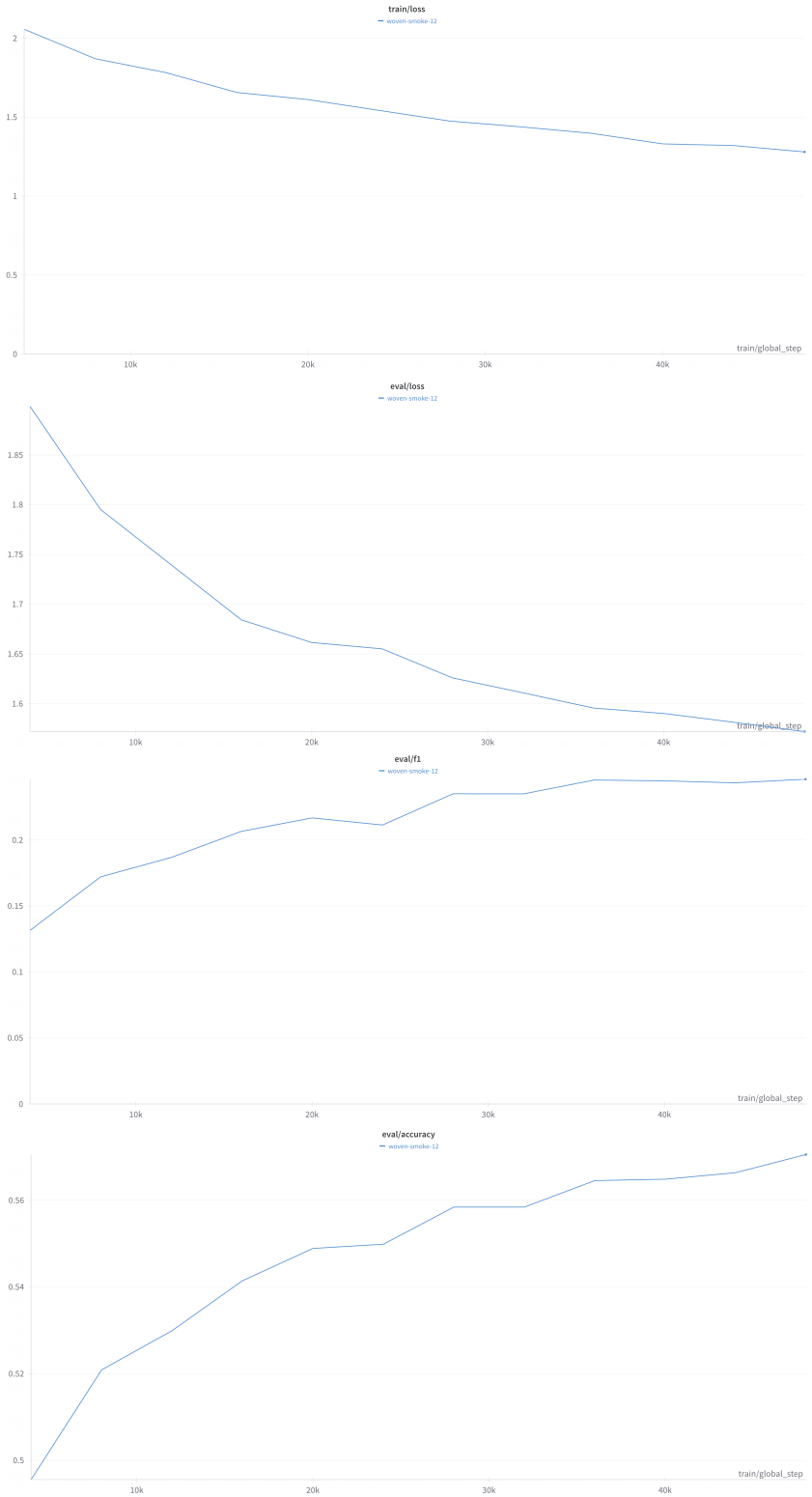

Figure 1: From top to down: (a) Fine-tuning loss, (b) Validation loss, (c) F1 score on QQP test set, (d) Accuracy on MNLI test set of a single random seed (seed 42) of RoBERTa-base as it undergoes fine-tuning.

Table 1: $|\Delta$ LMC$|$ metrics between the two random seeds between start and end of QQP fine-tuning

Metric	BERT-base-uncased	RoBERTa-base	RoBERTa-large
Barrier Height (BH)	0.0044	0.0119	0.0132
Convexity Gap (CG)	0.0135	0.0258	0.0371
Improved CG (ICG)	0.0228	0.0497	**0.0557**

Table 2: $|\Delta$ LMC$|$ metrics between the two random seeds between start and end of MNLI fine-tuning

Metric	BERT-base-uncased	RoBERTa-base	RoBERTa-large
Barrier Height (BH)	0.0445	0.0490	0.0511
Convexity Gap (CG)	0.4939	0.6003	**0.6011**
Improved CG (ICG)	0.3881	0.5917	0.5939

- We consistently found that across all models, for both QQP and MNLI fine-tuning, there is a noticeable jump when going from BERT-base to RoBERTa-base model. As the ΔLMC (all negative) increases when going from BERT to RoBERTa-base to RoBERTa-large, this supports our hypothesis that higher pretraining of base models, for a fixed amount of fine-tuning, leads to producing more linearly connected models.
- As each value in 1 and 2 was negative, this supports our second hypothesis well. Namely, this indicates that since the LMC metrics converge closer to 0 between start and end of fine-tuning for any pre-trained models with fixed weights, this means that for a given pre-trained model, more fine-tuning on a task complementary to the pre-training task results in more linearly connected models.

6 Analysis

While working through the experiments, we often found that improved convexity gap (ICG) resulted in higher values than the other two LMC metrics for any fixed amount of fine-tuning on a particular base model. While this was not the case in every single instance, this may indicate that the focus on sub-segment-maximization and future recursive metrics could result in displaying a sharper and more accurate drop in path dependence between random seeds. Below, we categorize some broad qualitative results using the observations we made in the previous section.

- In text classification models of BERT and RoBERTa's size, we can be reasonably confident that we obtain decreasing values of CG and ICG over the course of fine-tuning on NLI tasks, indicating that qualitatively, the model is pushed further down a particular basin during the fine-tuning process. This also results in strongly linearly connected random seeds for the pretrained models.
- We validate our second hypothesis, and our results indicate that for a given amount of fine-tuning time on NLI tasks, **models that have undergone higher pre-training will emerge as more LMC** than models that have not been trained past convergence. Although further work on a higher number of checkpoints and larger models could help establish more confidence in these results, our results display a clear indication of this trend.

7 Conclusion

In this work, by considering text classification fine-tuning on pre-trained models of variable sizes, we found that (a) more fine-tuning results in higher linearly connected models, pushing the network deeper down its loss basin, and (b) larger models display a higher amount of linear connectivity, and for fixed amounts of fine-tuning produce highly LMC random seeds.

Validating these results for a larger number of random seeds, a wider collection of datasets, and performing the analysis for models of larger size are natural future steps to further establish a grounding in this work.

References

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. 2018. Essentially no barriers in neural network energy landscape.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2022. Linear connectivity reveals generalization strategies.

Victor Lecomte, Kushal Thaman, Rylan Schaeffer, Naomi Bashkansky, Trevor Chow, and Sanmi Koyejo. 2023. What causes polysemanticity? an alternative origin story of mixed selectivity from incidental causes.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.