Linear Connectivity Reveals Generalization Strategies

Jeevesh Juneja [1] **Rachit Bansal** [1] **Kyunghyun Cho** [2] **João Sedoc** [2] **Naomi Saphra** [2]

Abstract

It is widely accepted in the mode connectivity literature that when two neural networks are trained similarly on the same data, they are connected by a path through parameter space over which test set accuracy is maintained. Under some circumstances, including transfer learning from pretrained models, these paths are presumed to be linear. In contrast to existing results, we find that among text classifiers (trained on MNLI, QQP, and CoLA), some pairs of finetuned models have large barriers of increasing loss on the linear paths between them. On each task, we find distinct clusters of models which are linearly connected on the test loss surface, but are disconnected from models outside the cluster—models that occupy separate basins on the surface. By measuring performance on existing diagnostic datasets, we find that these clusters correspond to different generalization strategies: one cluster behaves like a bag of words model under domain shift, while another cluster uses syntactic heuristics. Our work demonstrates how the geometry of the loss surface can guide models towards different heuristic functions.

1. Introduction

Modern training methods are capable of discovering high-performance parameters for neural networks on a variety of tasks. Although models trained with similar procedures tend to exhibit similar in-domain (ID) performance on these tasks, they exhibit diverse decision boundaries (Benton et al., 2021). In particular, models with similar performance often differ when presented with examples that fall far from the training data manifold (Somepalli et al., 2022).

In NLP, generalization behavior is usually characterized structurally through the use of diagnostic challenge sets. Previous studies of model behavior on out-of-distribution (OOD) linguistic structures show that finetuned models can exhibit variation in compositional generalization and performance on challenge sets (McCoy et al., 2020; Zhou et al., 2020). For example, in natural language inference tasks, some models seem to deploy strategies during OOD generalization that incorporate no position information at all (McCoy et al., 2019). To the best of our knowledge, these different generalization behaviors have never been linked to the geometry of the loss surface. In order to explore how barriers in the loss surface expose a model's generalization strategy, we consider the case of text classification. We focus in particular on Natural Language Inference (NLI; Williams et al., 2018; Consortium et al., 1996), as well as paraphrase and grammatical acceptability tasks.

We find that NLI models tend to rely on one of two strategies, both of which exhibit similar loss on the test set. We characterize these strategies as, roughly, **syntax-aware** and **syntax-unaware**. They fall into two respective basins; the syntax-aware basin contains functions that tend to rely on heuristics around the behavior of constituents, while functions in the syntax-unaware basin rely on lexical bag-of-words heuristics. We find that in NLI and paraphrase tasks, models that perform similarly on the same challenge sets are linearly connected without barriers on the ID loss surface, but they tend to be disconnected from models with different generalization behavior. Our code and models are public.[1]. Our main contributions are:

- We develop a metric based on linear mode connectivity, the **convexity gap**, and an accompanying method for clustering models into basins. In contrast with existing work in computer vision (Neyshabur et al., 2020), we find that transfer learning can lead to different basins over different finetuning runs.
- We align the basins to specific generalization behaviors. In NLI, they correspond to a preference for either syntactic or lexical overlap heuristics. On a paraphrase task, basins likewise split on behavior under word order permutation.
- We confirm that these basins trap a portion of finetuning

[1]Delhi Technological University [2]New York University. Correspondence to: Jeevesh Juneja <jeeveshjuneja@gmail.com>, Naomi Saphra <nsaphra@nyu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Code: https://github.com/aNOnWhyMooS/connectivity; Models: https://huggingface.co/connectivity

runs, which become increasingly disconnected from the other models as they train. Based on this behavior, it may be possible to predict heuristics from early connectivity.

2. Identifying generalization strategies

Finetuning on standard GLUE (Wang et al., 2018) datasets often leads to models that perform similarly on in-domain (ID) test sets (Sellam et al., 2021). To evaluate the functional differences between these models, we need to evaluate their generalization to diagnostic datasets. In this paper, we study the variation of performance on these existing diagnostic datasets. We will call models with poor performance on the generalization set **heuristic models** while those with high performance will be **generalizing models**. We study three tasks with diagnostic sets: MNLI (Williams et al., 2018) with the diagnostic set HANS (McCoy et al., 2019), QQP (Wang et al., 2017) with PAWS-QQP (Zhang et al., 2019), and CoLA (Warstadt et al., 2018) with CoLA-OOD (the last task is in Appendix A).

Natural Language Inference NLI is a common testbed for NLP models. This binary classification task poses a challenge in modeling both syntax and semantics. The input to an NLI model is a pair of sentences such as:

- *Premise:* The dog scared the cat.
- *Hypothesis:* The cat was scared by the dog.

Here, the label is positive or **entailment**, because the hypothesis can be inferred from the premise. If the hypothesis were, "The dog was scared by the cat", the label would be negative or **non-entailment**. We use the MNLI (Williams et al., 2018) corpus, and inspect loss surfaces on the "matched" validation set.[2]

NLI models often "cheat" by relying on heuristics, such as overlap between either individual lexical items or syntactic constituents shared by the premise and hypothesis. If a model relies on lexical overlap, both the entailed and non-entailed examples above might be given positive labels, because all three sentences contain "scared", "dog", and "cat". McCoy et al. (2019) responded to these shortcuts by creating HANS, a challenge set of sentence pairs that violate such heuristics:

- **Lexical overlap (HANS-LO):** Entails any hypothesis containing the same words as the premise.
- **Subsequence:** Entails any hypothesis containing contiguous sequences of words from the premise.
- **Constituent:** Entails any hypothesis containing syntactic subtrees from the premise.

[2]An unmatched validation set is also available, which includes different sources and topics from the training set. The test set labels are not public for MNLI or QQP, so we use the validation set only.

Unless otherwise specified, we use the non-entailing HANS subsets for measuring reliance on heuristics, so higher accuracy on HANS-LO indicates less reliance on lexical overlap.

Paraphrase Quora Question Pairs (QQP; Wang et al., 2017) is a common paraphrase corpus. We use the PAWS-QQP (Zhang et al., 2019) dataset as the diagnostic set for identifying different generalization behaviors learnt by the models. PAWS-QQP contains QQP sentence pairs in which the words have been permuted in order to construct pairs that can mean different things, even though the set of tokens remains the same. In other words, PAWS-QQP contains pairs that may violate a lexical overlap heuristic.

2.1. Experimental details

All models are initialized from `bert-base-uncased` with a linear classification head and finetuned using Google's default hyperparameters. MNLI models are those provided by McCoy et al. (2020).

3. Linear Mode Connectivity

Models discovered by SGD are generally connected by paths over which the loss is maintained (Draxler et al., 2019; Garipov et al., 2018). If we limit such paths to be linear, however, connectivity is no longer guaranteed. We may still find, however, that two parameter settings θ_A and θ_B, which achieve equal loss, can be connected by linear interpolation (Nagarajan and Kolter, 2019; Goodfellow et al., 2015) without any increase in loss. In other words, loss $\mathcal{L}(\theta_\alpha; X_{\text{train}}, Y_{\text{train}}) \leq \mathcal{L}(\theta_A; X_{\text{train}}, Y_{\text{train}})$ and $\mathcal{L}(\theta_\alpha; X_{\text{train}}, Y_{\text{train}}) \leq \mathcal{L}(\theta_B; X_{\text{train}}, Y_{\text{train}})$ in each parameter setting θ_α defined by a scalar $0 \leq \alpha \leq 1$:

$$\theta_\alpha = \alpha\theta_A + (1-\alpha)\theta_B \tag{1}$$

A number of results suggest that high performance is closely tied to the linear mode connectivity of the models in question (Frankle et al., 2020; Entezari et al., 2021; Neyshabur et al., 2020). Our results complicate the narrative around linear mode connectivity (Fig. 1). While we find that models with high performance on OOD data were indeed linearly connected to each other, models with low performance were also linearly connected to each other. It seems that the heuristic and generalizing models occupy two different linear basins, with barriers in the ID loss surface between models in each of these two basins.

HANS performance during interpolation: It is clear from Fig. 1 that, when interpolating between LO-heuristic models, HANS-LO loss significantly improves further from the end points. This finding implies that the heuristic basin does contain more syntax-aware models. In contrast, the

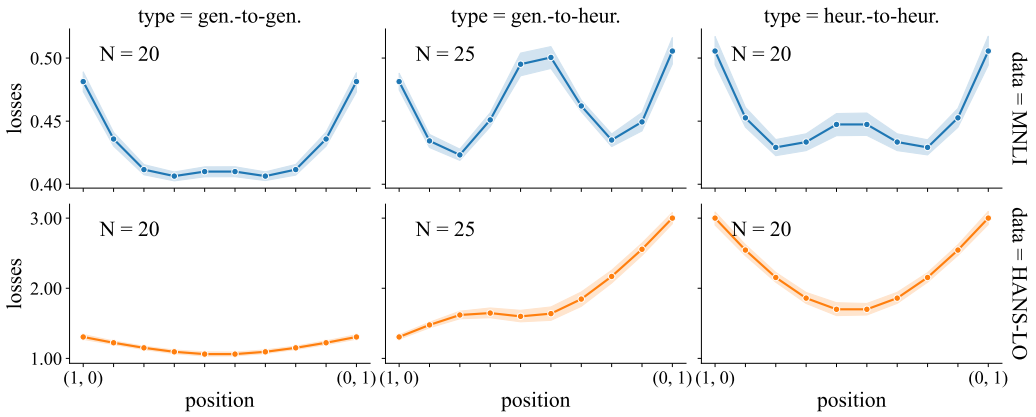


Figure 1. Loss during linear interpolation between pairs of models taken from the 5 top (gen.) and 5 bottom (heur.) in HANS-LO accuracy. Heuristic models tend to be poorly connected to the rest of the surface, although they are well connected to each other. N indicates number of pairs. Position on the x-axis indicates the value of α during interpolation.

syntax-aware basin shows only a very slight improvement in heuristic loss during interpolation, even though the improvement in broad coverage loss is more substantial than in the position-unaware basin.

Connections over 2 dimensions: To understand the loss topography better, we present planar views of these models using code[3] from Benton et al. (2021). We see how, in the plane covering the heuristic and generalizing models, a central barrier intrudes on the linear connecting edge between heuristic and generalizing models (Fig. 2). On the other hand, the heuristic and generalizing models each occupy separate planes that exhibit only a small central barrier. Visibly, this barrier is smallest for the perimeter composed of generalizing models and largest for the mixed perimeter.

4. The Convexity Gap

One possibility we argue against is that the increasing loss between models with different heuristics is actually an effect of sharper minima in the heuristic case. There is a significant body of work on the controversial (Dinh et al., 2017) association between wider minima and generalization in models (Li et al., 2018; Keskar et al., 2017; Hochreiter and Schmidhuber, 1997; Huang et al., 2020).

Instead, we find that clusters of linearly connected models are far better predictors of generalization behavior than optimum width is. To identify such clusters, we define a metric based on linear mode connectivity, the **convexity gap** (**CG**), and use it to perform spectral clustering.

Entezari et al. (2021) define a barrier's height on a linear path from θ_1 to θ_2 as (for $\alpha \in [0, 1]$):

$$\mathbf{BH}(\theta_1, \theta_2) = \sup_{\alpha}[\mathcal{L}(\alpha\theta_1 + (1-\alpha)\theta_2) - (\alpha\mathcal{L}(\theta_1) + (1-\alpha)\mathcal{L}(\theta_2))] \quad (2)$$

We define the convexity gap on a linear path from θ_1 to θ_2 as the maximum possible barrier height on any sub-segment of the linear path joining θ_1 and θ_2. With $\gamma, \beta \in [0, 1]$,

$$\mathbf{CG}(\theta_1, \theta_2) = \sup_{\gamma,\beta} \mathbf{BH}(\gamma\theta_1 + (1-\gamma)\theta_2, \beta\theta_1 + (1-\beta)\theta_2) \quad (3)$$

Our metric exposes stronger clustering patterns than either area under the curve (AUC) or **BH** (evidence in Appendix E).

4.1. Clustering

The basins that form from this **CG** distance metric are visible based on connected sets of models in the distance heatmap (Fig. 4a). To quantify basin membership into a prediction of HANS performance, we perform spectral clustering, with the distances between points defined as **CG**-distance on the ID loss surface. Using the difference between distances from each cluster centroid $\|w - c_1\|$ and $\|w - c_2\|$, we see a significantly larger correlation with HANS performance (Fig. 3a), compared to a baseline of the model's ϵ-sharpness (Fig. 3b). Furthermore, a linear regression model offers significantly better performance based on spectral clustering membership than on sharpness.

In Fig. 4b, we see the heuristic that defines the larger cluster: constituent overlap. Models that perform well on constituent overlap diagnostic sets tend to fall in the basin containing models biased towards lexical overlap. This behavior characterizes the two basins: the larger basin is syntax-aware

[3] https://github.com/g-benton/loss-surface-simplexes

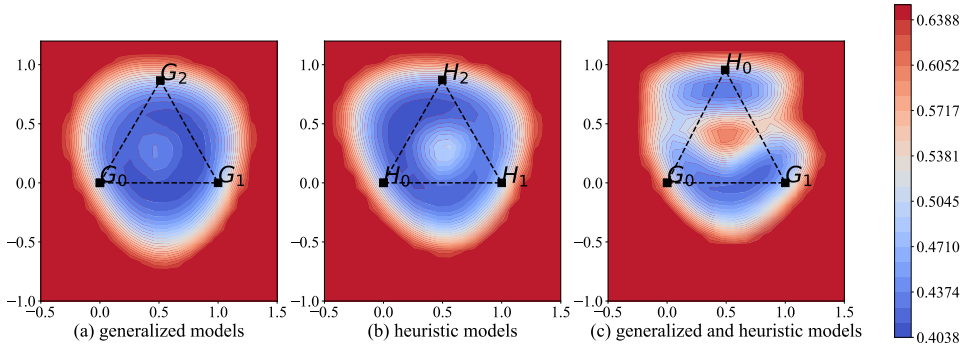
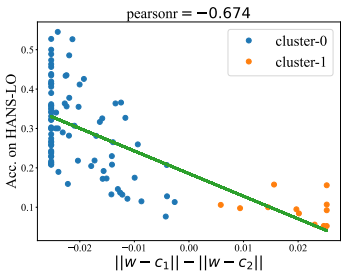
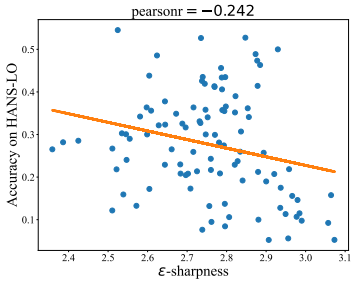


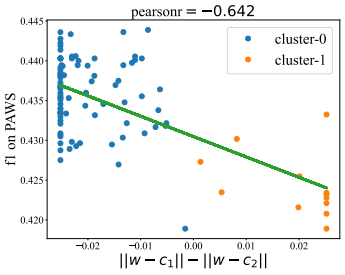
Figure 2. Planar views of simplices connecting NLI models in the MNLI matched validation set loss surface. The points $G_{0\ldots2}$ and $H_{0\ldots2}$ denote the generalized and heuristic models respectively.



(a) HANS-LO performance vs difference between distances from the cluster centroids c_1, c_2 in the **CG**-based spectral clustering space.

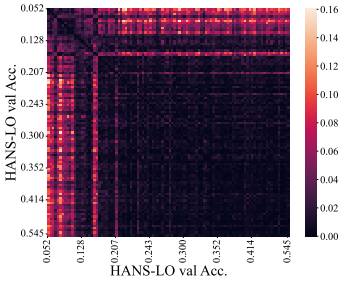


(b) HANS-LO performance vs ϵ-sharpness.

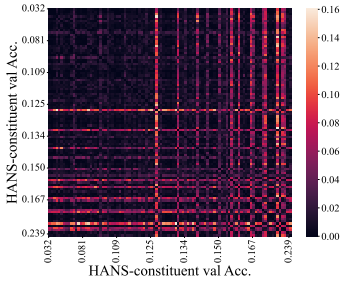


(c) PAWS-QQP performace vs centroid distance.

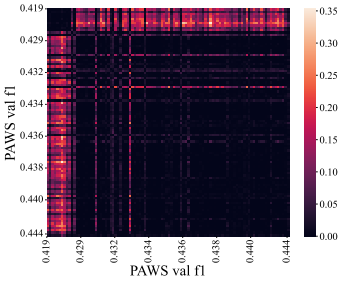
Figure 3. Using features of the ID loss landscape to predict reliance on OOD heuristics. Least squares fit shown.



(a) NLI models sorted by increasing accuracy on HANS-LO.



(b) NLI models sorted by increasing accuracy on HANS-constituent.



(c) QQP models models sorted according to f1 score on PAWS-QQP.

Figure 4. Color indicates **CG** distance between the x- and y-axis models.

(tending to acquire heuristics that require awareness of constituent structure), while the smaller basin is syntax-unaware (acquiring heuristics that rely only on unordered sets of words).[4]

4.1.1. QQP

On QQP, we similarly find strong differentiation between distinct clusters of linearly connected models (Fig. 4c). As in our procedure on NLI, we study the distance from the centroids formed by spectral clustering with **CG** defining the distance metric, and find cluster membership to be a strong predictor of generalization on PAWS-QQP (Fig. 3c).

4.2. Generalization basins trap training trajectories

At this point we have aligned different basins with particular generalization behaviors, but a skeptic may propose that the smaller cluster contains models which have simply not trained long enough. Under this conjecture, continuing to train them would eventually place them in the larger, syntax-aware cluster. We find that this is not the case, as shown in Fig. 5. Marginal members of the smaller cluster, those which fall closer to the cluster boundary, may drift towards the larger cluster later in training. However, models that are more central to the cluster actually become increasingly solidified in their cluster membership later in training.

These results have two important implications. First, they confirm that these basins constitute distinct local minima which can trap optimization trajectories. Additionally, our findings suggest that early on in training, we can detect whether a final model will follow a particular heuristic. The latter conclusion is crucial for any future work towards practical methods of directing or detecting desirable generalization strategies during training (Jastrzebski et al., 2021).

5. Conclusions

In one view of transfer learning, a pretrained model may select a prior over generalization strategies by favoring a particular basin. Neyshabur et al. (2020) found that models initialized from the same pretrained weights are linearly connected, suggesting that basin selection is a key component of transfer learning. We find that a pretrained model may not commit exclusively to a single basin, but instead favor a small set of them. Furthermore, we find that linear connectivity can indicate shared generalization strategies under domain shift, as evidenced by results on NLI, paraphrase, and linguistic acceptability tasks.

The split between generalization strategies can potentially explain results from the bimodality of CoLA models (Mosbach et al., 2021) to wide variance on NLI diagnostic sets (McCoy et al., 2020). Because weight averaging can find parameter settings that fall on a barrier, we may even explain why weight averaging, which tends to perform well on vision tasks, fails in text classifiers (Wortsman et al., 2022). Future work that distinguishes generalization strategy basins could improve the performance of such weight ensembling methods.

[4]It is not enough to declare one basin to be merely *position* aware, as reliance on subsequence overlap is a poor predictor of basin membership (Appendix D).

References

G. W. Benton, W. J. Maddox, S. Lotfi, and A. G. Wilson. Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. *arXiv:2102.13042 [cs, stat]*, Feb. 2021. URL http://arxiv.org/abs/2102.13042. arXiv: 2102.13042.

T. F. Consortium, R. Cooper, D. Crouch, J. V. Eijck, C. Fox, J. V. Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, S. Pulman, T. Briscoe, H. Maier, and K. Konrad. Using the framework, 1996.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp Minima Can Generalize For Deep Nets. *arXiv:1703.04933 [cs]*, May 2017. URL http://arxiv.org/abs/1703.04933. arXiv: 1703.04933.

F. Draxler, K. Veschgini, M. Salmhofer, and F. A. Hamprecht. Essentially No Barriers in Neural Network Energy Landscape. *arXiv:1803.00885 [cs, stat]*, Feb. 2019. URL http://arxiv.org/abs/1803.00885. arXiv: 1803.00885.

R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. *arXiv:2110.06296 [cs]*, Oct. 2021. URL http://arxiv.org/abs/2110.06296. arXiv: 2110.06296.

J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3259–3269. PMLR, Nov. 2020. URL https://proceedings.mlr.press/v119/frankle20a.html. ISSN: 2640-3498.

T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *arXiv:1802.10026 [cs, stat]*, Oct. 2018. URL http://arxiv.org/abs/1802.10026. arXiv: 1802.10026.

I. J. Goodfellow, O. Vinyals, and A. M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv:1412.6544 [cs, stat]*, May 2015. URL http://arxiv.org/abs/1412.6544. arXiv: 1412.6544.

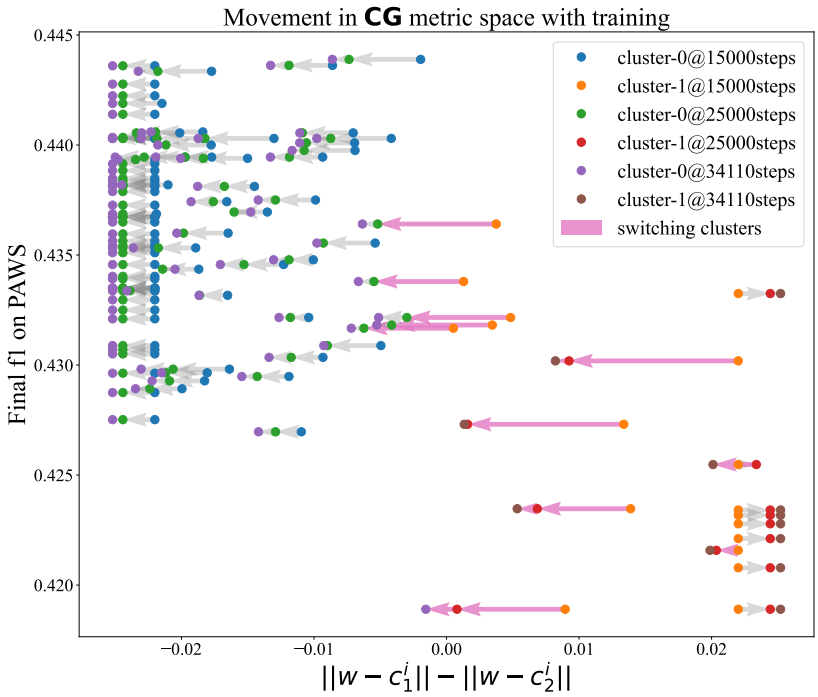


Figure 5. Movement between clusters during finetuning. Syntax-unaware models near the cluster boundary move towards the syntax-aware cluster, but central models in both clusters move closer to their respective centroids.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9:1–42, 1997.

W. R. Huang, Z. Emam, M. Goldblum, L. Fowl, J. K. Terry, F. Huang, and T. Goldstein. Understanding Generalization through Visualizations. *arXiv:1906.03291 [cs, stat]*, Nov. 2020. URL http://arxiv.org/abs/1906.03291. arXiv: 1906.03291.

S. Jastrzebski, D. Arpit, O. Astrand, G. B. Kerg, H. Wang, C. Xiong, R. Socher, K. Cho, and K. J. Geras. Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4772–4784. PMLR, July 2021. URL https://proceedings.mlr.press/v139/jastrzebski21a.html. ISSN: 2640-3498.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ArXiv*, abs/1609.04836, 2017.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the Loss Landscape of Neural Nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6389–6399. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-net.pdf.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

R. T. McCoy, J. Min, and T. Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *ArXiv*, abs/1911.02969, 2020.

T. McCoy, E. Pavlick, and T. Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

M. Mosbach, M. Andriushchenko, and D. Klakow. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv:2006.04884 [cs, stat]*,

Mar. 2021. URL http://arxiv.org/abs/2006.04884. arXiv: 2006.04884.

V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/05e97c207235d63ceb1db43c60db7bbb-Abstract.html.

B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 512–523. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html.

T. Sellam, S. Yadlowsky, I. Tenney, J. Wei, N. Saphra, A. D'Amour, T. Linzen, J. Bastings, I. R. Turc, J. Eisenstein, D. Das, and E. Pavlick. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *ICLR*, Sept. 2021. URL https://openreview.net/forum?id=K0E_F0gFDgA.

G. Somepalli, L. Fowl, A. Bansal, P. Yeh-Chiang, Y. Dar, R. Baraniuk, M. Goldblum, and T. Goldstein. Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective. *arXiv:2203.08124 [cs]*, Mar. 2022. URL http://arxiv.org/abs/2203.08124. arXiv: 2203.08124.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *ArXiv*, abs/1804.07461, 2018.

Z. Wang, W. Hamza, and R. Florian. Bilateral multi-perspective matching for natural language sentences. *ArXiv*, abs/1702.03814, 2017.

A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv:2203.05482 [cs]*, Mar. 2022. URL http://arxiv.org/abs/2203.05482. arXiv: 2203.05482.

Y. Zhang, J. Baldridge, and L. He. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*, 2019.

X. Zhou, Y. Nie, H. Tan, and M. Bansal. The Curse of Performance Instability in Analysis Datasets: Consequences, Source, and Suggestions. *arXiv:2004.13606 [cs]*, Nov. 2020. URL http://arxiv.org/abs/2004.13606. arXiv: 2004.13606.

A. Linguistic Acceptability

The Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018) is a set of acceptable and unacceptable English sentences collected from the linguistics literature. Linguistics has a longstanding practice of studying minimal changes that render sentences ungrammatical; one CoLA example is the pair of sentences "Betsy buttered the toast" (acceptable) and "Betsy buttered at the toast" (unacceptable).

CoLA includes an ID val/test set, where the examples are taken from the same linguistics papers that the training set uses. However, it also includes an OOD diagnostic val/test set. The diagnostic sets are taken from a different set of linguistics papers, so in order for a model to perform well on CoLA-OOD, it must transfer a general ability to recognize unacceptable English sentences, rather than simply learning the set of acceptability rules described in the ID sources.

A.1. Experimental details

We found that default settings on the HuggingFace (Wolf et al., 2020) training script[5] resulted in more pronounced barriers between models, compared to the Google script we used for NLI and QQP.[6] Because our goal is to study the relationship between barriers and generalization, we therefore chose to use Huggingface for our CoLA analysis. Like in our other experiments, we kept the default hyperparameters, which differ slightly from the Google script. The CoLA models were trained for 6 epochs with a learning rate of 2×10^{-5}, a batch size of 32 samples, and no weight decay. This script uses the AdamW(Loshchilov and Hutter, 2017) optimizer too, with a linear learning rate decay schedule but no warm-up.

A.2. Clustering

In CoLA, there are very few barriers between finetuned models. A single model out of the 48 finetuned accounted for all substantial interpolation convexity gaps (Fig. 6a), thus forming its own one-point cluster when using **CG** as a distance metric for spectral clustering. This outlier model outperformed all others on OOD generalization (Fig. 6b), suggesting that CoLA is another task where models with different generalization behavior are disconnected.

[5] https://github.com/huggingface/transformers/blob/main/examples/flax/text-classification/run_flax_glue.py

[6] The major difference between scripts appears to be the lack of different initializations of classification head between huggingface runs. That is, different data order is the only source of SGD noise in HuggingFace runs. However, it is likely that the presence or absence of a second cluster during the sweep is due to random chance, given that we see only a single model out of 48 falling into the outlier cluster in these results.

B. A Notion of Convex Basins

From the interpolation plots(Figures 1, 2), it seems that the loss surface over low dimensional subspaces(accessible via SGD) of parameter space is composed of multiple, approximately convex, valleys. To formalize this, we define a notion of relaxed convex basin in the parameter space.

Defintion .1. *For resolution $\epsilon \geq 0$, we define an **ϵ-convex basin** as a convex set S, such that, for any set of points $w_1, w_2, ..., w_k \in S$ and any set of coefficients $\alpha_1 \dots \alpha_n \geq 0$ where $\sum_k \alpha_k = 1$, a relaxed form of Jensen's inequality holds:*

$$\mathcal{L}(\sum_{k=1}^{n} \alpha_k w_k) \leq \epsilon + \sum_{k=1}^{n} \alpha_k \mathcal{L}(w_k) \tag{4}$$

C. Theoretical result on convexity gaps

Theorem C.1. *An ϵ-convex basin will have $\mathbf{CG}(w_1, w_2) \leq \epsilon$ for every pair of models w_1, w_2 on its surface.*

Proof. Recall the definition of convexity gap as the maximum value of the barrier height of any segment θ_1, θ_2 along the interpolation between w_1 and w_2 from Equation 3:

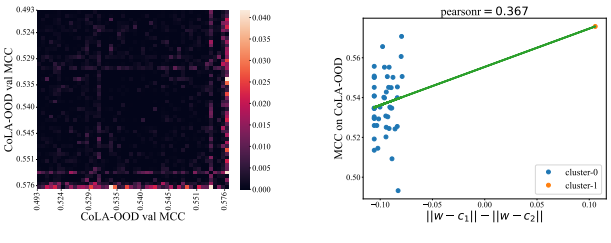
$$\mathbf{CG}(w_1, w_2) = \sup_{\gamma,\beta \in [0,1]} \mathbf{BH}(\gamma w_1 + (1-\gamma)w_2, \beta w_1 + (1-\beta)w_2) \tag{5}$$

As we are in an ϵ-basin, the defining inequality from Equation 4 holds $\forall \theta_1, \theta_2 \in \epsilon-$convex basin :

$$\mathcal{L}(\sum_{k=1}^{n} \alpha_k \theta_k) \leq \epsilon + \sum_{k=1}^{n} \alpha_k \mathcal{L}(\theta_k) \tag{6}$$

Applying this for $n = 2$:

$$\mathcal{L}(\alpha_1\theta_1 + (1-\alpha_1)\theta_2) - (\alpha_1\mathcal{L}(\theta_1) + (1-\alpha_1)\mathcal{L}(\theta_2)) \leq \epsilon \quad \forall \theta_1, \theta_2 \in \epsilon\text{-basin} \tag{7}$$



(a) (b)

Figure 6. (a) **CG** heatmap on CoLA models, sorted by OOD validation. (b) A scatter-plot of cluster membership versus performance on CoLA-OOD dev set.

Hence the supremum of the quantity on LHS is also $\leq \epsilon$. Seeing the definition of **BH** from Equation 2, we immediately see that:

$$\mathbf{BH}(\theta_1, \theta_2) \leq \epsilon \qquad \forall \theta_1, \theta_2 \in \epsilon\text{-basin} \tag{8}$$

As **CG** is the supremum of **BH** over elements within the $\epsilon-$convex basin only, we have:

$$\mathbf{CG}(w_1, w_2) \leq \epsilon \qquad \forall w_1, w_2 \in \epsilon\text{-basin} \tag{9}$$

□

D. HANS performance on subsequence heuristics

Models that perform poorly on subsequence heuristics tend to fall in the bag-of-words basin. However, the clusters are less pronounced than for either constituent or lexical overlap heuristics (Fig. 7).

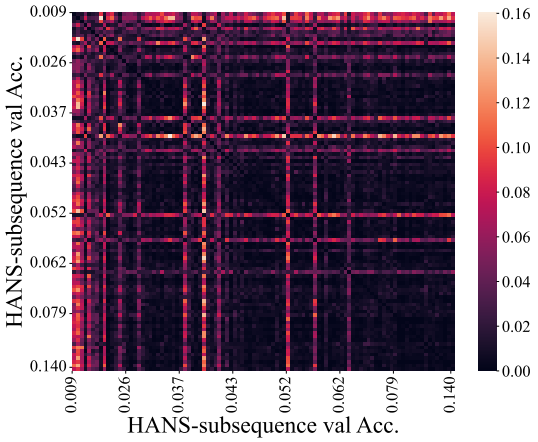


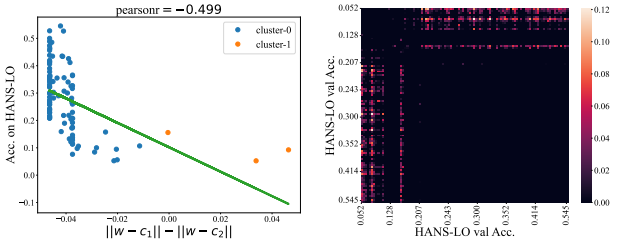
Figure 7. **CG** for model pairs, sorted by increasing performance on HANS-subsequence.

E. Alternative barrier measurements

In Fig. 8 we show NLI model clustering using the original barrier height metric (Equation 2) from Entezari et al. (2021). We can see that the Pearson's correlation coefficient is -0.49, which is far lower in magnitude than correlation in the **CG**-metric space.

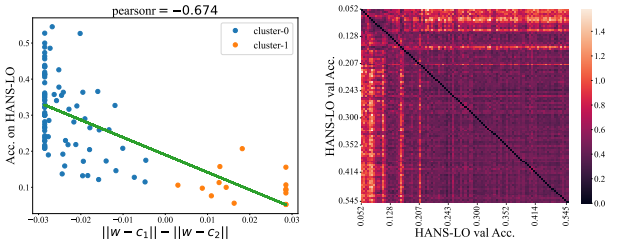
We also show the results for another metric, viz. area under the interpolation curve (AUC), in Fig. 9. Although this shows the same pearson's correlation coefficient, the clusters are much less crisp in the heatmap. In order to compute AUC, we add together the area between each point on the curve and the lowest point in the entire curve.

Finally, we consider the effect of Euclidean distance (Fig. 10). The clustering effect is extremely strong and



(a) (b)

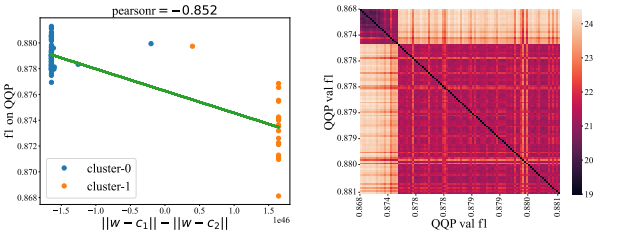
Figure 8. Relationship between **BH** on the MNLI validation loss surface and HANS-LO accuracy.



(a) (b)

Figure 9. Relationship between AUC on the MNLI validation loss surface and HANS-LO accuracy.

predictive of generalization behavior. It is worth considering the possibility that basins trap the models in a way that forces them to have larger Euclidean distances, but those Euclidean distances are the property that most determines the generalization strategy of the model.



(a) (b)

Figure 10. Relationship between Euclidean distance on the QQP validation loss surface and PAWS-QQP accuracy.