Using Local Complexity to Evaluate Out-of-Distribution Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite their growing ubiquity, the inner workings of deep neural networks are still largely a black box. Even in the case of classification tasks, common methods used to assess model performance do not give insight into whether the model will generalize to unseen data. In this extended abstract, we investigate local complexity (LC) Humayun et al. (2024b), a geometric measure of the input space, as a predictor of model performance on out-of-distribution (OOD) data. We find that LC alone is not sufficient to predict model generalization, but that it does capture meaningful information about the correctness of individual predictions, suggesting it may be useful as part of a larger set of tools to understand OOD generalization.

1 Introduction

Out-of-distribution (OOD) generalization–reliable performance when the deployment distribution differs from the training distribution–remains a central challenge in machine learning. Research addressing this problem ranges from developing algorithms to improve OOD generalization to identifying model properties that facilitate robust generalization, such as the stability of estimations under small data perturbations Gupta & Rothenhausler (2021).

Recently, Humayun et al. (2024b) introduced *local complexity (LC)*, a data-dependent geometric measure that approximates the local density of linear regions around inputs in networks with piecewise-linear activation functions. They proposed local complexity as a progress measure Barak et al. (2022) and linked decreases in LC in the final phase of training to grokking (delayed generalization) and increased adversarial robustness. Building on this work and motivated by the intuition that larger linear regions around the training data promote generalization, we investigate whether LC can serve as an effective predictor of model generalization capabilities.

The key question we ask is whether LC can predict OOD classification performance at the model or the per-example level. At the model level, we examine whether qualitative training-time LC trajectories predict performance on OOD data. We find that these dynamics alone are insufficient for reliable predictions about model generalization, primarily because LC dynamics are highly dependent on model architecture. However, at the per-example level, we find that LC is significantly lower for correctly classified OOD inputs compared to misclassified ones, suggesting that it captures a meaningful component of OOD generalization. This indicates that while LC may not work as a standalone predictor at the model level, it may complement other uncertainty measures, a direction we plan to explore in future work.

2 Local complexity

Local complexity, introduced by Humayun et al. (2024b), measures the density of spline partition regions that tile a deep neural network's (DNN) input space. DNNs map an input vector, x to an

Submitted to 1st Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

output vector y through a composition of affine and nonlinear functions. In particular, a network with activation function $\mathbf{a}$ and K layers can be written as

$$y = b_K + W_K \mathbf{a}\Big[b_{K-1} + W_{K-1}\mathbf{a}\big[\dots b_2 + W_2\mathbf{a}\left[b_1 + W_1\mathbf{a}\left[b_0 + W_0 x\right]\right]\dots\big]\Big]. \tag{1}$$

where b_k is the vector of biases for hidden layer $k+1$ and W_k is the weights matrix applied to the kth layer. In this work, the activation function $\mathbf{a}$ is always ReLU. Balestriero & Baraniuk (2018) showed that for any piecewise-linear activation function, (1) is a continuous piecewise-affine spline operator. That is, there is a partition Ω of the input space such that the network acts affinely on any region $\omega \in \Omega$. These are the spline partition regions whose density local complexity tracks.

Measuring local complexity For a convex region $\mathcal{V}$ in the input space of our network, local complexity can be computed in terms of the hyperplanes stemming from each neuron. For the kth layer of a network with weight matrix W_k, bias vector b_k, and output dimension d_k, the spline partition Ω_k of the input space to layer k can be written as the hyperplane arrangement where each hyperplane is associated to a neuron in layer k, that is, $\partial\Omega_k = \bigcup_{i=1}^{d_k} \mathcal{H}_k^{(i)}$, where $\mathcal{H}_k^{(i)} = \left\{x \in \mathbb{R}^{d_{k-1}} \colon \langle w_k^{(i)}, x\rangle + b_k^{(i)} = 0\right\}$, $w_k^{(i)}$ is the ith row of W_k, and $b_k^{(i)}$ is the ith entry of b_k.

To approximate the LC induced by the kth layer on $\mathcal{V}$, we simply count the number of regions in $\bigcup_i^{d_k} \Phi \cap \mathcal{H}_k^{(i)}$, where Φ is the embedded representation of $\mathcal{V}$ after being passed through layers 1 through $k-1$ of the network. To simplify computation, we consider the number of hyperplanes passing through Φ as a proxy for the LC of $\mathcal{V}$ at layer k. Following Humayun et al. (2024b) and utilizing their code base[1], to measure how local complexity changes throughout training, we randomly sample data points and construct randomly-oriented, P-dimensional ℓ_1-norm balls with radius r centered at each data point. We then count the number of hyperplanes passing through these neighborhoods to approximate the local complexity in that area for a given layer. In this work, we take $P = 2$ and $r = 0.5$. These choices are discussed further in Appendix B.2.

Training dynamics of local complexity In Section 4, we will compare the dynamics of local complexity across models. We use the word "dynamics" to refer to the general qualitative behavior, as well as phases described in Humayun et al. (2024b). In that work, the authors describe the *two descent phases* of LC. After initialization, they note the first descent. This phase does not always occur; it is dependent on the network parameterization and initialization. Then, in the ascent phase, region density accumulates around training and testing points until training interpolation is reached. Finally, in the second descent phase, also called the *region migration* phase, the nonlinearities shift towards the decision boundary leading to increased LC near the boundary and decreased LC away from the training data. They document these dynamics for a variety of model architectures and datasets. Additionally, the authors study how architecture and regularization influence LC dynamics in both Humayun et al. (2024b) and Humayun et al. (2023).

3 Experimental set up

To test model performance on out-of-distribution data, we train 25 different models on CIFAR-10[2] Krizhevsky et al. (2009). These models include architectures from the ResNet, DenseNet, and VGG families, among others. We trained models with and without data augmentation (random crop, random horizontal flip, and random erasing) to compare its effect on LC. A full list of models can be found in Appendix B.1, and details on model training in Appendix B.2. To simulate OOD data, we evaluated each of our models on a subset of images from CIFAR-10-Warehouse[3] (CIFAR-10-W) Sun et al. (2024), a collection of 180 datasets motivated by the observation that many collections of OOD testsets have a small number of domains or rely on synthetic corruptions. Specifically, we used the subset of CIFAR-10-W sourced from the internet image search engine, 360. This dataset has over 60,000 images separated into 12 different colors across the same classes as CIFAR-10. The images were chosen through keyword searches of the form "*color class*" (i.e., "red airplane").

[1] `https://github.com/AhmedImtiazPrio/grok-adversarial`; released with an MIT License

[2] `https://www.cs.toronto.edu/~kriz/cifar.html`; released with an MIT License

[3] Licensed under CC BY-NC 4.0 1

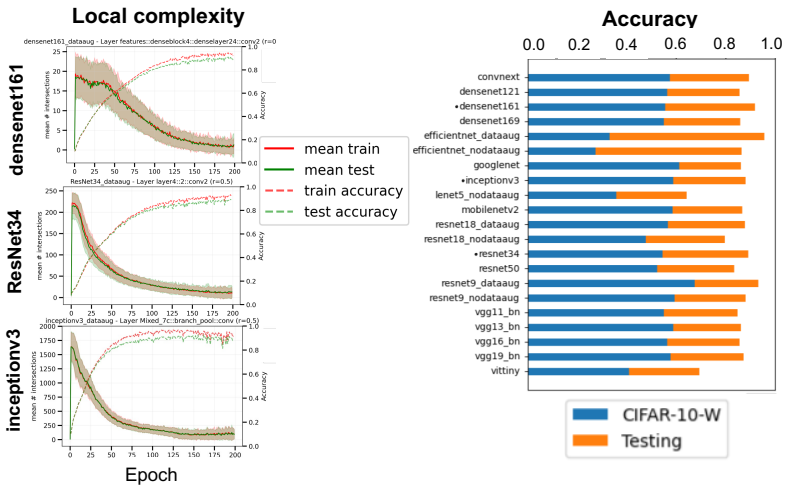

Figure 1: Left: LC on training and testing data (CIFAR-10) throughout training for three models. Differences in model architecture lead to different qualitative behavior in LC. Right: Accuracy on CIFAR-10-W is shown in blue, with accuracy on the CIFAR-10 test set overlaid in orange for comparison.

4 Results and Discussion

Different LC dynamics, similar OOD performance As in Humayun et al. (2023), our work corroborates that training-time LC trajectories depend on model architecture. Figure 1 shows three examples of LC trajectories, measured at the final layer before the classification layer. DenseNet-161 (left, top) is closest to having two descent phases[4]: after rising upon initialization, the mean number of regions intersected decreases slightly, and then flattens, before decreasing for the remainder of training. In contrast, for both ResNet-34 and InceptionV3 the sharp spike after initialization is followed by a much faster decrease. (Note that the maximum number of possible hyperplane intersections is directly related to the number of neurons, making it difficult to compare LC values directly between models with different architectures). Despite the qualitative differences in local complexity dynamics, DenseNet-161 and ResNet-34 perform very similarly on CIFAR-10-W, with 55.18% and 54.34% accuracy, respectively. While Inception-V3 and ResNet-34 have very similar LC dynamics, Inception-V3 outperforms ResNet-34, getting 58.41% accuracy. As shown in Figure 1, Inception-V3, DenseNet-161, and ResNet-34 are 87.74%, 91.44%, 88.83% accurate on the CIFAR-10 test set, respectively. Overall, the trends we saw in LC dynamics confirmed findings in Humayun et al. (2023) that LC is highly dependent on model architecture. (Figure 6 in Appendix E show the similarity in LC dynamics between models from the same architecture families.) Consequently, LC dynamics alone are not sufficient information to predict a model's ability to generalize to OOD data.

Comparing LC of correctly and incorrectly classified examples Though LC alone is not enough to predict model performance on OOD data overall, we found that it does reflect some information about the correctness of individual predictions. We evaluate DenseNet-161 and ResNet-34 on 128 random samples from each of the 12 color groups of CIFAR-10-W's 360 dataset, separate the data based on whether it was correctly or incorrectly classified by the model, and measure the difference in LC at the end of training (we omitted InceptionV3 because of its similarity in LC dynamics to ResNet-34). We find, on average, that the number of hyperplanes intersecting a neighborhood is higher among the incorrectly classified points than those correctly classified (Figure 3). In fact, computing t-tests on LC measurements by color and model reveals that the majority of these differences are statistically significant ($p \leq 0.05$). We found 19 of 24 t-tests were statistically significant. Further details can be found in Appendix C. A possible future direction is to train a classifier that considers LC, among other information, as input and predicts if an individual OOD example will be correctly classified.

[4]All three models use batch normalization, which could be the reason we see only a single descent phase.

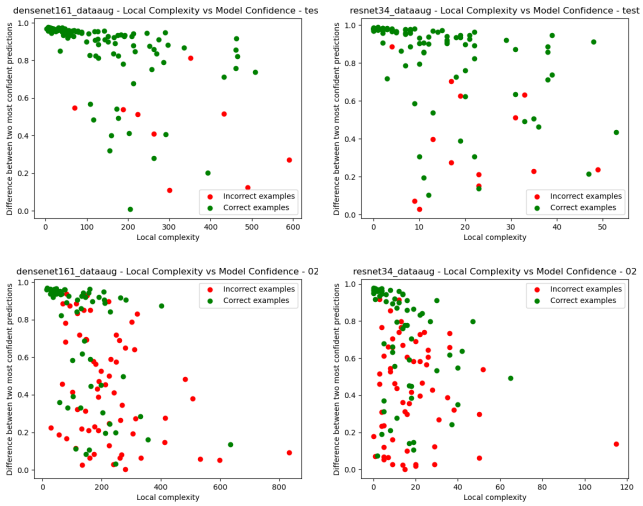

Figure 2: Top: Model confidence vs. local complexity for examples from the in-distribution CIFAR-10 holdout test set. To evaluate model confidence, we take the difference of the two highest softmax logits. Bottom: Model confidence vs. local complexity for examples from CIFAR-10-W in color '02' ("orange") on three models.

LC and model confidence Since average thresholded confidence is used to predict accuracy on unlabeled test data Garg et al. (2022), as an initial step towards understanding what features could be used in conjunction with LC, we study the relationship between LC and confidence. Our intuition is guided by the idea that the spline partitions will migrate towards the decision boundary throughout training Humayun et al. (2024b). So, we expect that an example with low local complexity will be firmly located within a particular label's region, implying that the model is confident in its prediction. Conversely, an example near the decision boundary will have high LC and low confidence. The plots of model confidence vs. LC on our in-distribution testing data shown in Figure 2, top, support this. We see that for both DenseNet-161 and ResNet-34 there are a cluster of correctly classified points in the upper left, while incorrectly classified points more often fall in the lower right.

Figure 2, bottom, shows model confidence vs. LC but for OOD data. In this case, we still see a distinct cluster of correctly classified points in the upper left, but also see more incorrectly classified examples in this area. This suggests that these examples fall firmly within the region of the input space for a particular label (far from the decision boundary), but that it is the incorrect label, reflecting one way in which data can be OOD. Additionally, we see many examples in the OOD data that fall in the lower left. These examples are not easily explained by our current understanding and warrant further investigation. One could use a tool like SplineCam Humayun et al. (2024a) to better understand where in the input space these OOD examples lie.

5 Future directions

The statistically significant difference in mean LC between correctly and incorrectly classified OOD samples indicates this measure captures meaningful aspects of OOD generalization and suggests several avenues for future study. Future work could use the full distribution of LC values rather than just the means, including class-specific patterns, to further understand model generalization and robustness. Notably, since this approach computes LC at the end of training, it can be extended to pretrained models, broadening its practical applications.

It would also be interesting to use LC to identify confusing training examples or important data features. Examples of questions include how removing high-LC training examples affects generalization performance, and whether analyzing which hyperplanes intersect neighborhoods most often could reveal key features. While the present work uses CIFAR-10 and CIFAR-10-W, in future work we plan to expand to other datasets and evaluate our hypotheses on a larger collections of models.

References

Balestriero, R. and Baraniuk, R. A spline theory of deep networks. In *International Conference on Machine Learning*. PMLR, 2018.

Barak, B., Edelman, B., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Garg, S., Balakrishnan, S., Lipton, Z., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Gupta, S. and Rothenhausler, D. The s-value: evaluating stability with respect to distributional shifts. In *Neural Information Processing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:234095751`.

Gupta, S. and Rothenhäusler, D. The s-value: evaluating stability with respect to distributional shifts. *Advances in Neural Information Processing Systems*, 36:72058–72070, 2023.

Hanin, B. and Rolnick, D. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*. PMLR, 2019.

Humayun, A. I., Balestriero, R., and Baraniuk, R. Training dynamics of deep network linear regions. *arXiv preprint arXiv:2310.12977*, 2023.

Humayun, A. I., Balestriero, R., Balakrishnan, G., and Baraniuk, R. Splinecam: Exact visualization and characterization of deep network geometry and decision boundaries. *arXiv preprint arXiv:2302.12828*, 2024a.

Humayun, A. I., Balestriero, R., and Baraniuk, R. Deep networks always grok and here is why. *arXiv preprint arXiv:2402.15555*, 2024b.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Mu, N. and Gilmer, J. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906:0.2337*, 2019.

Patel, N. and Montúfar, G. On the local complexity of linear regions in deep relu networks. *arXiv preprint arXiv:2412.18283*, 2024.

Sun, X., Xingjian, L., Wang, Z., Yang, Y., Huang, Z., and Zheng, L. Cifar-10-warehouse: Broad and more realistic testbeds in model generalization analysis. *arXiv preprint arXiv:2310.04414*, 2024.

Yu, H., Liu, J., Zhang, X., Wu, J., and Cui, P. A survey on evaluation of out-of-distribution generalization. *ArXiv*, abs/2403.01874, 2024. URL `https://api.semanticscholar.org/CorpusID:268248288`.

A Related work

OOD generalization A recent survey Yu et al. (2024) categorizes OOD evaluation into three groups based what test data it requires. OOD performance testing evaluates models when labeled test data is available, OOD performance prediction evaluates models when unlabeled test data is available, and OOD intrinsic property characterization aims to discover properties of models that inform OOD generalization when no test data is available. Examples of intrinsic properties include characteristics like stability of estimates under small perturbations Gupta & Rothenhäusler (2023) and flatness Foret et al. (2020). When unlabeled test data is available, Garg et al. (2022) proposes the method Average Thresholded Confidence to predict accuracy on OOD data using model confidence.

Local complexity Patel & Montúfar (2024) uses a slightly different definition of local complexity and develops theory that explains some of the results in Humayun et al. (2024b). Namely, they show that their formulation of local complexity is an upper bound on the total variation of the network over the input space. They also connect local complexity to local rank, the average dimension of the feature manifold at intermediate layers. Though they do not discuss local complexity directly, Hanin & Rolnick (2019) studies spline partitions and investigates alternative ways to quantify the changing partition regions.

B Additional details on experimental setup

B.1 Complete list of trained models

We trained 25 models on CIFAR-10. Unless stated otherwise, all models were trained with data augmentation. Below is the complete list:

- ConvNeXt
- DenseNet-121
- DenseNet-161
- DenseNet-169
- EfficientNet
- EfficientNet without data aug.
- GoogLeNet
- Inception-V3
- LeNet-5 without data aug.
- MobileNet-V2
- ResNet-18
- ResNet-18 without data aug.
- ResNet-34
- ResNet-50
- ResNet-9
- ResNet-9 without data aug.
- VGG-11
- VGG-11 with batch norm.
- VGG-13
- VGG-13 with batch norm.
- VGG-16
- VGG-16 with batch norm.
- VGG-19
- VGG-19 with batch norm.
- ViTTiny

Many of these models were listed here[5] as suggestions for use on CIFAR-10. Of these models, we chose to exclude VGG-11, VGG-13, VGG-16, and VGG-19 without batch normalization from analysis because they never learned better than random chance.

B.2 Experimental choices

Local complexity hyperparameters We made choices in our experimental design based on observations of preliminary MNIST experiments. For example, we set the radius $r = 0.5$ for the ℓ_1-neighborhoods (Section 2). We experimented with various sizes of radii and settled on 0.5 as it seemed to capture the most change in LC–with larger r, the number of intersections was always quite high and smaller caused the neighborhoods to be too small to consistently intersect any of the hyperplanes.

[5] `https://zenodo.org/badge/latestdoi/195914773`

We chose the dimension of the neighborhoods to be $P = 2$. Our experiments with MNIST showed that increasing the dimension increased the scale of the number of intersections per neighborhood, but did not tend to change the overall dynamics. Thus, we chose the minimum dimension for computational efficiency.

Dataset selection We chose to use CIFAR-10-W over a dataset with synthetic corruptions to better simulate real-world encounters of unseen data. We chose to experiment on only a subset of CIFAR-10-W. There are many different datasets from various internet search engines within CIFAR-10-W. In addition, for some search engines, they create cartoon datasets by searching "*color class* cartoon" to further push the data out of distribution. The dataset also includes images generated using diffusion models. Within CIFAR-10-W, we focused on the search engine 360 for simplicity and the existence of an analogous cartoon version, though we ultimately did not analyze that data.

Model confidence In Section 4, we evaluate model confidence using the difference of the two highest softmax logits. We chose this computation instead of simply taking the highest value as it suggests the model prefers a single label over all others which we interpret as a data point lying far from the decision boundary.

Training details When training each of our models, we used stochastic gradient descent as our optimizer and trained for 200 epochs. We set weight decay to be 0.01, learning rate to be 0.1, momentum to be 0.9, and used a batch size of 128. We used a scheduler to set the learning rate to follow a linear warmup schedule followed by a cosine annealing schedule. We used the default train/test split of CIFAR-10 with 50,000 training points and 10,000 test points. All models were trained on a single NVIDIA A100.

C Full t-test results

We chose to run independent t-tests as we wish to compare the means of a statistic between two different populations. We sample 128 points randomly from CIFAR-10-W and assume that the LC of each point is an independent observation. We find that the correctly and incorrectly classified examples have similar variances. The table below shows the p-values for each of the t-tests described in Section 4. The bold values are statistically significant ($p \leq 0.05$).

Color	DenseNet-161	ResNet-34
Red '01'	**0.01688**	**0.00176**
Orange '02'	**0.00021**	0.07751
Yellow '03'	**3.05161 e-07**	**1.30424 e-05**
Green '04'	**0.00059**	**5.04387 e-06**
Light Blue '05'	**0.00812**	0.27444
Blue '06'	**4.41172 e-09**	**0.00011**
Purple '07'	**0.00019**	0.25146
Pink '08'	**1.0775 e-06**	**8.61807 e-05**
Brown '09'	**0.00041**	**9.86942 e-05**
Gray '10'	**2.17557 e-06**	**7.99038 e-08**
White '11'	**0.04719**	**6.84716 e-05**
Black '12'	0.69873	0.08998

In addition to studying differences between means, we also examined how LC changes throughout training, both overall (Figure 3) and for individual examples (Figure 4). Although there is notable overlap between the correctly and incorrectly classified examples, the statistically significant difference between the two distributions suggests that we may be able to discern which examples will be correctly classified using local complexity.

D Preliminary MNIST experiments

Preliminary experiments on MNIST[6] LeCun et al. (2010) informed our approach to studying CIFAR-10. We trained 10 different models on MNIST for 10 epochs each and computed local complexity

[6]Licensed under MIT License

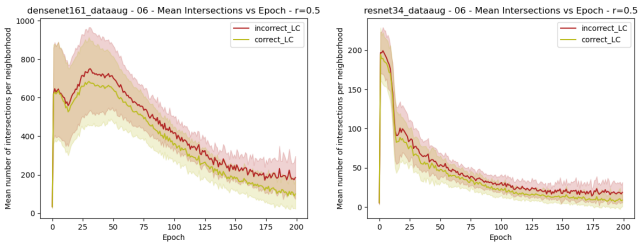

Figure 3: Mean intersections per neighborhood for examples from CIFAR-10-W in color '06' ("blue") from three models. Examples are separated based on if they were correctly (green) or incorrectly (red) classified by the model. The mean among the incorrectly classified examples is higher than among the correctly classified examples. This is true across all colors. The statistical significance is discussed in Section 4.

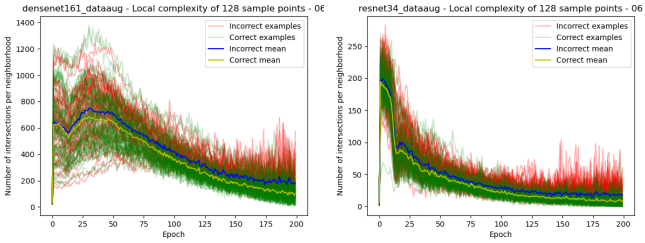

Figure 4: Number of intersections per neighborhood for 128 examples taken from CIFAR-10-W in color '06' ("blue"). Examples that are correctly classified are shown in green and those incorrectly classified in red. This allows us to see how local complexity changes throughout training for each example individually.

throughout training on 100 training and 100 testing examples. We then evaluated each of the models on MNIST-C[7] Mu & Gilmer (2019), a corrupted, synthetic version of MNIST. We compared accuracy and LC across models and across the 15 corruptions in MNIST-C. We found that even when a model was able to achieve high accuracy on corrupted data, the LC dynamics for "brightness" and "fog" were qualitatively different than for other corruptions. These two corruptions are the only two that edit the contrast of the original images leading us to wonder if contrast is a particularly important feature in the model's decision-making process. Further study could reveal if local complexity can identify how influential certain data features are in model predictions.

E Additional figures

[7]Licensed under Apache License Version 2.0

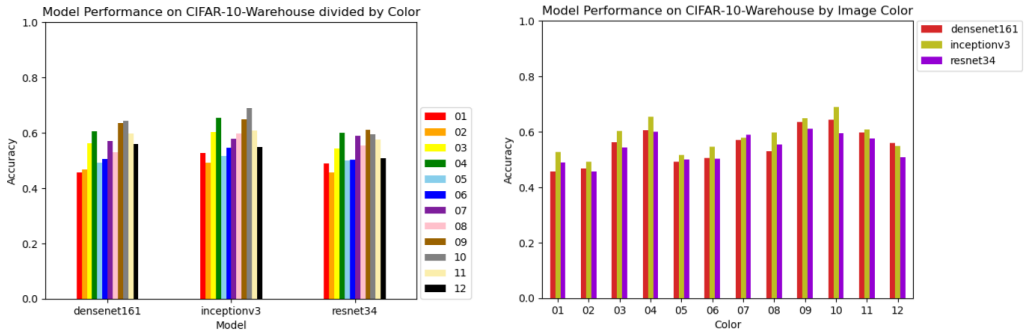

Figure 5: Two bar plots showing model performance separated by each of the color groups within CIFAR-10-W. Left: comparing the performance between the 12 colors by each of the 3 models. Right: comparing performance between models on each of the 12 colors.

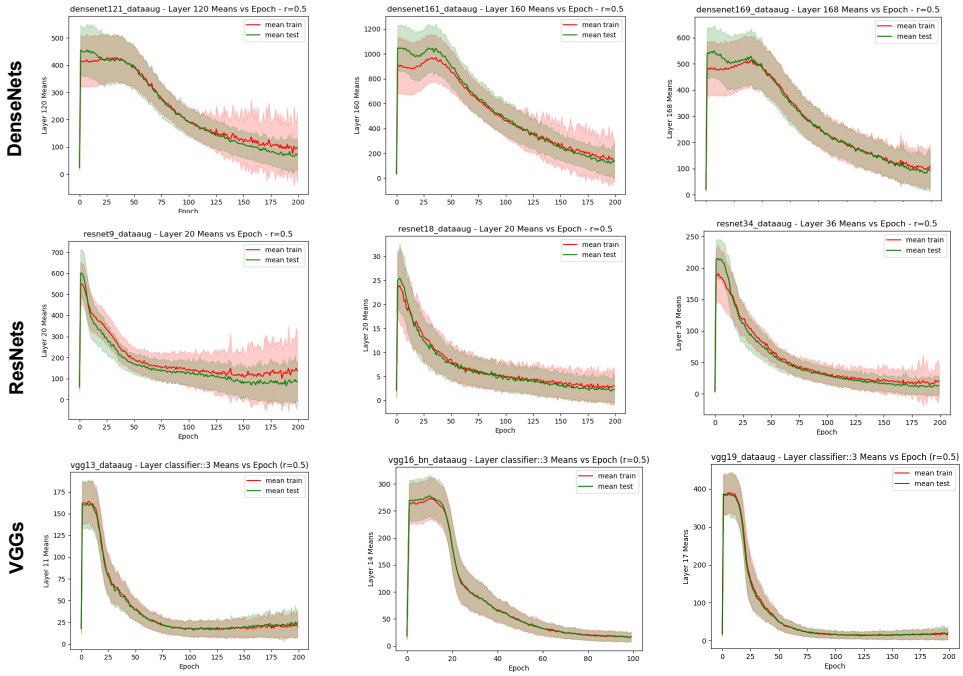

Figure 6: LC at the final layer before the classification layer for three DenseNet architectures (DenseNet-121, DenseNet-161, DenseNet-169), three ResNet architectures (Resnet-9, ResNet-18, ResNet-34) and three VGG architectures (VGG13, VGG16, and VGG19). LC dynamics look very similar between models with similar architecture.

TAG-DS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the findings and motivations of the paper. We discuss our main claim: that local complexity captures a meaningful component of OOD generalization, but is alone not sufficient to predict OOD generalization.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We are clear in our explanation that this is work performed on a single dataset and provides encouraging evidence that local complexity may be a helpful tool to understand OOD generalization. We make no claims that our trends necessarily hold in general and discuss making these findings more robust in Section 5. We perform t-tests in Section 4 and in the Appendix (Section C) the circumstances that lead us to believe this is an accurate test to use.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not contain any theoretical results. We make only empirical claims about the relationship between local complexity and OOD generalization.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of our set-up in the paper. An overview can be found in Section 3 with further details in the Appendix (Sections B.1, D, and B.2)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This work is part of an ongoing project. We intend to make our code available when we submit the full-length version of this paper.

   Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, this information can be found in Section 3 with further details in the Appendix (Sections B.1, D, B.2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: On figures where we report a mean, we provide a shaded region around the mean representing one standard deviation above and below the mean. We compute t-tests in Section 4 to support our claim that there is a significant difference in the mean LC between correctly and incorrectly classified examples. Further details of these t-tests, including all p-values are provided in Section C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, this information can be found in Appendix B.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work generally seeks to improve our ability to predict OOD generalization. We claim to have made partial progress towards this goal, but do not entirely solve it. As such, we see no immediate societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks. We exclusively used pre-existing models and datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are cited in the bibliography. When appropriate, URLs are provided. We were unable to find the license for CIFAR-10, but it is linked and properly cited. All other assets have their licenses provided in footnotes when they are first mentioned.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not include crowdsourcing nor human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method of development does not involve LLMs as any important components. We consulted LLMs only for suggestions on minor tasks (i.e., formatting figures in matplotlib), not for any scientific portion of our research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.