Benchmark Hill Climbing During Large Model Pretraining: Some Preliminary Investigations

Rajan Vivek
Stanford University

Kawin Ethayarajh
Stanford University
(Advisor)

Diyi Yang
Stanford University
(Advisor)

Douwe Kiela
Stanford University
(Advisor)

Abstract

Scaling up machine learning model size, pretraining data, and pretraining time has resulted in incredible leaps in model performance across a breadth of language and vision tasks. However, relying on this strategy alone to improve downstream performance is computationally expensive and ignorant of other factors that impact performance including pretraining data and model hyperparameters. Here we ask: how can we predict a model's downstream performance on a specific benchmark– composed of multiple disparate tasks– at pretraining time in a lightweight manner? We propose a broad framework for rapidly predicting benchmark performance via zero-shot performance on a small number of downstream examples (selected in a manner that minimizes redundancy). We investigate a few stepping stones towards achieving this vision including assessing the correlation between zero-shot and fine-tuned performance, as well as measuring data relatedness at both the task and data point level through analysis of training dynamics. We then propose a range of next steps.

1 Motivation

The incredible success of large pretrained models has marked a paradigm shift in machine learning from training single-task models to fine-tuning preexisting models for a downstream use case. Downstream use cases may be composed of multiple distinct tasks, for example a conversational system that needs to understand both language sentiment and entailment. While scaling up data, model size, and pretraining time tends to improve downstream performance, this strategy alone is computationally expensive and often myopic. It ignores other factors that strongly impact downstream performance including model architecture, pretraining data, pretraining self-supervision method, and performance saturation [1,2]. While precisely defining the phenomena underlying these interactions is a crucial research direction, here we ask a practical question: How can we predict downstream performance on a specific benchmark at pre-training time in a lightweight manner? Such a technique would allow practitioners to identify the optimal model to fine-tune for their application and allow researchers to optimize the pretraining of their own models. This would accelerate both research of large pretrained models and their adoption in real-world systems, while also minimizing environmental impact.

2 Related Work

Work quite similar in spirit to our own is that of [1], which performs a large-scale investigation into predicting downstream task accuracy using upstream (pretraining) accuracy. Across thousands of experiments, they discover a nonlinear relationship reliably modeled with a power law curve, in which increasing upstream accuracy results in a saturation of downstream accuracy at a value dependent on the relatedness of the pretraining and downstream tasks. The curve appears for both fine-tuned downstream accuracy (where all layers are fine-tuned on 1000 downstream examples)

and few-shot downstream accuracy (where all layers are frozen except a linear head trained on 1-25 downstream examples). We note that this work does not consider 1) the relationship between few-shot and fine-tuned performance, 2) the optimal strategy for selecting representative downstream test examples, and 3) the case where the downstream task is composed of multiple disparate tasks.

The concept of selecting a small number of representative points from a dataset is well-explored in active learning and coreset literature. These strategies fall into two broad categories: uncertainty-based sampling and density-based sampling [3]. The former characterizes a dataset by selecting points close to decision boundaries, where a model tends to be uncertain. The latter characterizes a dataset by selecting points that evenly span the feature distribution such that all points are well-represented. However, [4] emphasize that selecting points for evaluation rather than learning warrants distinct selection criteria: 1) unlearnable points should still be selected, 2) selection bias should be removed (this bias tends to be helpful in active learning [5]), and 3) the variance of the evaluation loss estimate should be minimized. They propose a method of selecting high loss points using a surrogate model, leveraging the unbiased estimator for evaluation performance R_{PURE} from [5], which achieves minimal variance when the highest loss points are selected. We note that existing active selection literature does not consider the case where the dataset is composed of multiple disparate tasks.

[6] propose a method for characterizing datasets based on the training dynamics of individual points. Plotting the model's confidence in the true class and variability in confidence for each point throughout training reveals 3 distinct regions: ambiguous regions (which are important for generalization), easy-to-learn regions (which are important for convergence), and hard-to-learn regions (which tend to correspond to label errors). This work inspires our own investigations into correlations within training dynamics as a metric for data point relatedness.

3 Methods

Algorithm 1 Benchmark Performance Prediction via Zero-Shotting Selected Points

$D_{benchmark} \leftarrow \{\{$ example $x_{1,i}$, label $y_{1,i}\}_{i=1}^{L_1} ... \{$ example $x_{N,i}$, label $y_{N,i}\}_{i=1}^{L_N}\}$
$B \leftarrow$ budget ▷ maximum number of points to evaluate
$Models \leftarrow$ candidate models
$C \leftarrow$ CORRELATE($D_{benchmark}$) ▷ inter-task and inter-point correlations
$D_{test} \leftarrow$ SELECT($D_{benchmark}$, B, C)
$M_{BEST}, ZS_{BEST} \leftarrow$ None, -Inf ▷ best model and zero shot performance
for $M \in Models$ **do**
  $ZS \leftarrow$ LogLikelihood(M(D_{test}[x]), D_{test}[y])
  if $ZS > ZS_{BEST}$ **then**
    $M_{BEST}, ZS_{BEST} \leftarrow M, ZS$
  end if
end for

3.1 Formulation

We aim to develop a procedure for benchmark hill climbing that follows the general recipe shown in Algorithm 1. In short, we intend to develop a benchmark data point selection strategy such that the zero shot performance of a model on these points is predictive of the fine-tuned model performance on that benchmark. Considering that the benchmark may be composed of multiple tasks with disparate difficulties and levels of relatedness, we intend to incorporate dataset and data point relatedness metrics to find a minimal subset that represents the benchmark well. We perform investigations toward three initial objectives:

1. Demonstrate that the zero shot performance of a pretrained model on a downstream task correlates well with its fine-tuned performance.
2. Identify a strategy for evaluating the relatedness of tasks.
3. Identify a strategy for evaluating the relatedness of points, both within and across tasks.

These objectives are stepping stones toward developing a selection strategy corresponding to SELECT in Algorithm 1. The candidate models may correspond to various model architectures, models

pretrained on different corpora, or various checkpoints of a single model on a specific pretraining corpus.

3.2 Objective 1: Demonstrate Correlation Between Zero Shot and Fine-Tuned Performance

We train a 12-layer BERT model on a 100M Wikipedia corpus until convergence (300k steps, 15 epochs). Every 10000 steps, we fine-tune the checkpoint on 1000 examples of SST-2 and evaluate on 400 examples. Only the last transformer layer and a classifier head are trained during fine-tuning. Additionally, we evaluate zero-shot performance using the prompts shown in Figure 1. We measure the correlation between the fine-tuned accuracy and zero shot accuracy for each prompt.

3.3 Objective 2: Evaluate the Relatedness of Tasks with Training Dynamics

Evaluating the relatedness of tasks is important for determining how to partition our evaluation budget across different tasks. Specifically, we want to know what information is gained about a model's performance on one task from knowing a model's performance on another. We experiment with quantifying task relatedness by measuring correlations between zero shot performances on various tasks across model checkpoints. Intuitively, we expect that tasks with similar performance trends across many models are likely to be related. To test this, we correlate the confidence in the correct label of 29 checkpoints of our 12-layer BERT model across two movie review classification tasks from distinct sources (SST-2 and IMDB), 1 language proficiency classification task (One Stop English), and 1 news classification task (AG News). We evaluate on 50 randomly-selected examples from each task. (Classes were roughly balanced for each). We hypothesize that the movie review tasks will correlate positively with each other, while the language proficiency task will not correlate strongly with the other tasks (due to its emphasis on language semantics over content).

3.4 Objective 3: Evaluating the Relatedness of Points with Training Dynamics, both within and across Tasks

Evaluating the relatedness of points is important for avoiding redundancy in the points we select to evaluate. We consider a brute force $O(n^2)$ approach (where n is the number of points) in which the correlation in model confidence in the correct class between all points is calculated in a pairwise fashion. We also consider a further optimized $O(n)$ approach in which a high degree polynomial is fit to the training dynamics of each point and the polynomial coefficients are clustered to identify related points. We consider the correlation between these two approaches, as well as the optimal degree polynomial to use.

4 Results and Discussion

4.1 Correlation Between Zero-Shot and Fine-Tuned Performance

Figure 1 shows the zero shot accuracy on SST-2 of 6 different prompts across training. Figure 2 shows the correlations between these trends and fine-tuned performance on SST-2 (where only the last layer and classifier head were trained on 1000 examples from SST-2). We see that multiple prompts are positively correlated, though only one prompt has a correlation with a p-value less than 0.01. Further analysis across additional tasks, models, and prompts is required to assess how these findings generalize and to identify a prompt selection strategy. (Issues with the compute cluster delayed these experiments).

4.2 Evaluating Task Relatedness with Training Dynamics

Figure 3 shows the average negative log likelihood training dynamics of four tasks throughout training of our 12-layer BERT. These results were selected from the best performing Prompt Source prompt for each task. Note that a lower negative log likelihood indicates greater confidence in the correct class. For most tasks, zero shot performance peaks well before the last checkpoint.

Figure 4 shows the correlations between these trends. As expected, SST-2 and IMDB are strongly positively correlated with statistical significance. We also see that One Stop English is negatively correlated with all other datasets (though this is only statistically signficant for SST-2), perhaps because this is the only task that emphasizes language semantics rather than content. These results suggest that training dynamics may be informative for determining task similarity, though other techniques should be explored to identify the most robust solution.

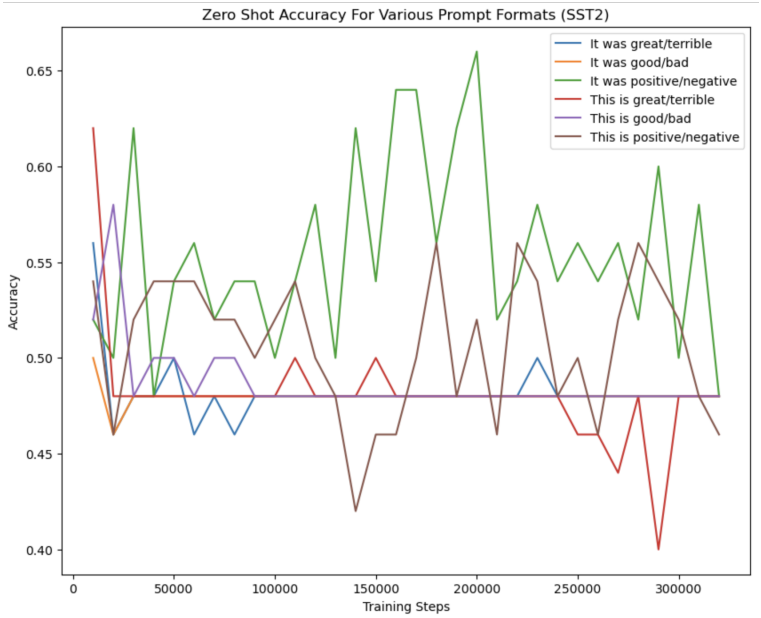


Figure 1: Zero Shot Accuracy on SST-2 by Prompt.

Prompt Format	**Correlation** w/ Fine-tuned Performance (p-value)
It was great/terrible	**0.264** (0.1)
It was good/bad	**0.021** (0.9)
It was positive/negative	**-0.184** (0.27)
This is great/terrible	**0.159** (0.34)
This is good/bad	**0.502** (0.001)
This is positive/negative	**0.30** (0.066)

Figure 2: Correlation between SST-2 Fine-Tuned Performance and Zero-Shot Performance by Prompt. Red indicates p-value < 0.01

4.3 Evaluating Data Point Relatedness with Training Dynamics

We next extend our training dynamics correlation analysis to the data point level. We first compute negative log likelihood trend correlations between all pairs of points within each dataset. Multi-dimensional scaling is used to convert this similarity information to a continuous space (Figure 5). We observe that this results in clear clustering between class labels. This result holds for multi-class

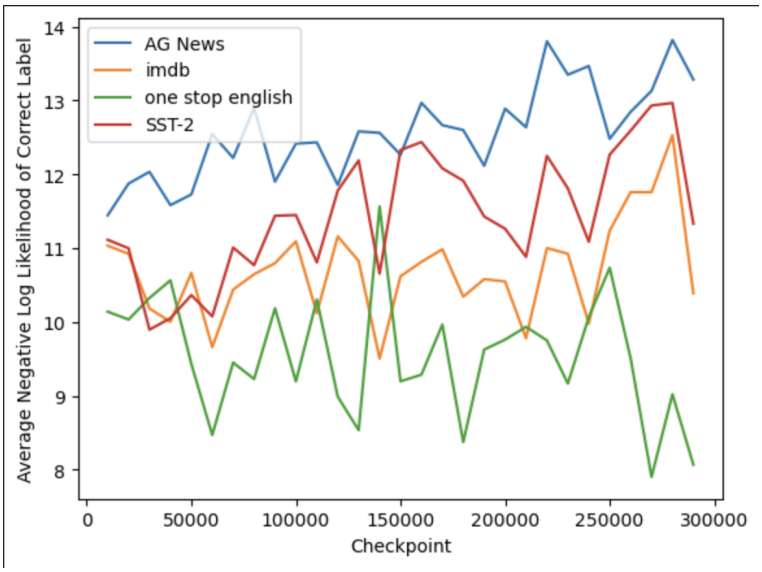


Figure 3: Zero Shot Average Negative Log Likelihood of Correct Class Throughout Training for Various Tasks. (Lower is better).

Dataset	SST-2	IMDB	One Stop English	AG News
SST-2	1	X	X	X
IMDB	**0.77** (1E-6)	1	X	X
One Stop English	**-0.39** (0.037)	**-0.32** (0.088)	1	X
AG News	**0.51** (0.004)	**0.27** (0.164)	**-0.33** (0.08)	1

Figure 4: Correlation Matrix of Training Dynamics from Different Tasks (Refers to Figure 3). P-values are shown in parentheses. Red indicates a p-value less than 0.01 and green indicates a p-value less than 0.05.

tasks as well (Figure 6). There does not appear to be strong structure within any given class, though we note that the relative class diversity can be determined by the density of the corresponding cluster (e.g. see 'World' vs 'Business' in Figure 6b.

In Figure 7, the same technique is applied to pairs of datasets. We see that the negative classes of SST-2 and IMDB correlate well, as do the positive classes. Figure 7b scales this analysis to seven classes across two datasets, though no significant insights are found.

In Figure 8, we measure the correlation between the $O(n)$ polynomial interpolation technique described in Section 3.4 and the pairwise $O(n^2)$ training dynamic correlation technique discussed thus far. We see that for 3 of the 4 datasets evaluated, the two techniques correlate strongly with a Pearson correlation coefficient around 0.9. The ideal polynomial seems to be approximately the lowest degree polynomial that perfectly interpolates the points (with degree 1 less than the number of points). Polynomials with degrees immediately above or far below this amount tend to have poorer correlation. Unfortunately this technique does not seem to always perform strongly, only achieving a correlation coefficient around 0.6 for the AG News task.

5 Conclusion and Next Steps

We propose a high-level framework for predicting downstream benchmark performance of a model at pretraining time by zero-shot evaluating selected downstream points. We find that zero-shot

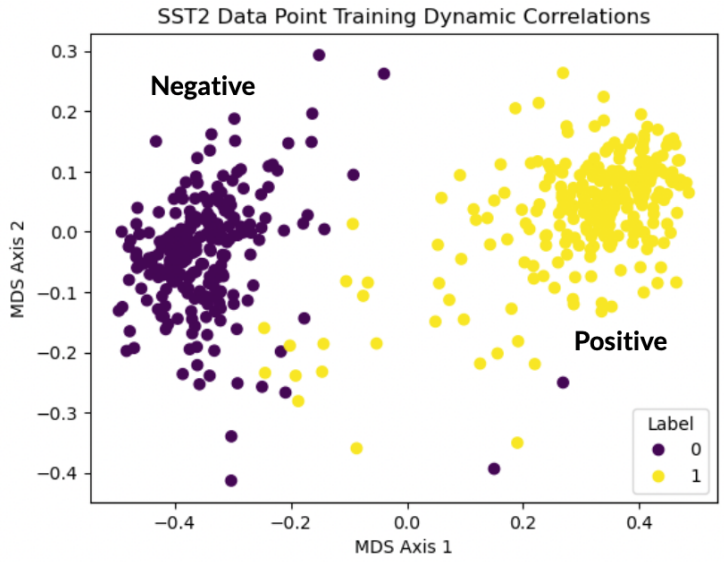
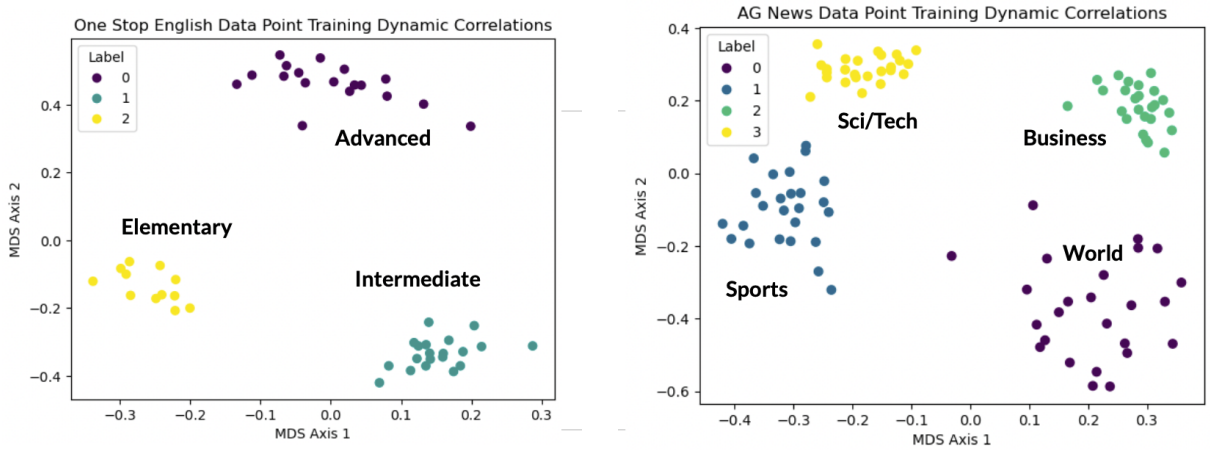


Figure 5: Multi-dimensional Scaling of SST-2 Data Point Training Dynamic Correlations.



(a) One Stop English

(b) AG News

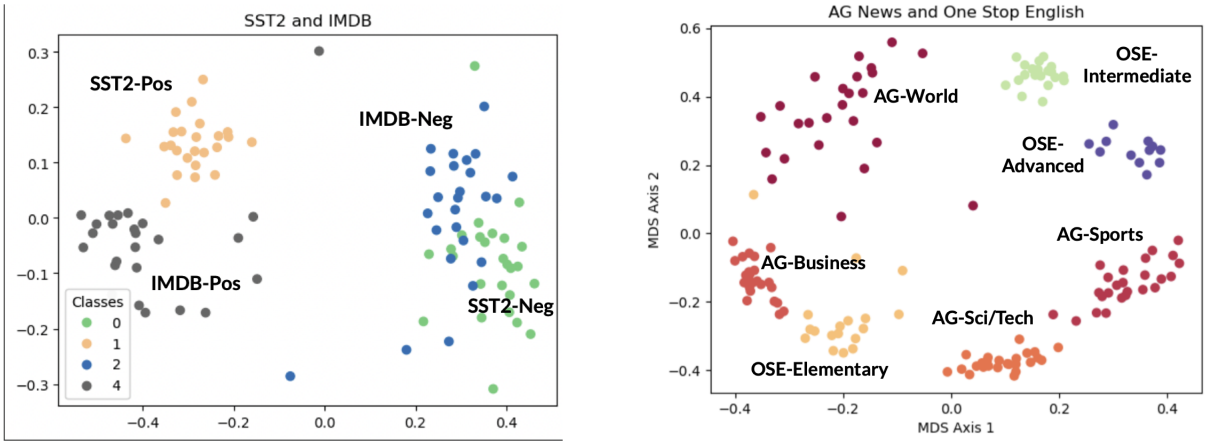
Figure 6: Multi-dimensional Scaling of Multiclass Task Data Point Training Dynamic Correlations.

prompting demonstrates positive correlation with fine-tuned performance, depending on the choice of prompt. Additionally, we show that correlating the training dynamics of different tasks matches human intuition about task relatedness. This technique can be extended to the data point level and made more efficient through fitting polynomials and clustering their coefficients.

We have multiple next steps in mind. We plan to scale our analysis of fine-tuning and zero-shot correlation to more models, tasks, and prompts to establish greater empirical justification and strategy for selecting prompts. We note that correlations in training dynamics may only be an approximate measure of relatedness and intend to explore more theoretically-driven techniques including that of [7], which computes task embeddings based on estimates of the Fisher information matrix associated with network parameters. We then plan to mathematically formalize the data point selection process and ideally identify cheap heuristics that approximate the optimal solution well.

6 References

[1] Abnar, Samira, et al. "Exploring the limits of large scale pre-training." *arXiv preprint arXiv:2110.02095* (2021).



(a) SST-2 and IMDB

(b) One Stop English and AG News

Figure 7: Multi-dimensional Scaling of Data Point Training Dynamic Correlations Across Tasks.

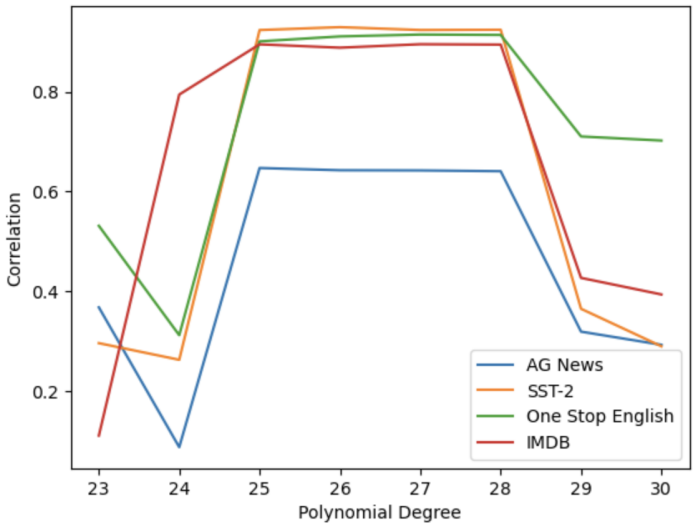


Figure 8: Correlation Between Polynomial Coefficient Euclidean Distance and Pointwise Training Dynamic Correlation. Note that 29 checkpoints are used.

[2] Anonymous Authors. "The Role of Pretraining Data in Transfer Learning." *Under submission at ICLR 2023.*

[3] Emam, Zeyad Ali Sami, et al. "Active Learning at the ImageNet Scale." *arXiv preprint arXiv:2111.12880* (2021).

[4] Kossen, Jannik, et al. "Active testing: Sample-efficient model evaluation." International Conference on Machine Learning. PMLR, 2021.

[5] Farquhar, Sebastian, Yarin Gal, and Tom Rainforth. "On statistical bias in active learning: How and when to fix it." *arXiv preprint arXiv:2101.11665* (2021).

[6] Swayamdipta, Swabha, et al. "Dataset cartography: Mapping and diagnosing datasets with training dynamics." arXiv preprint arXiv:2009.10795 (2020).

[7] Achille, Alessandro, et al. "Task2vec: Task embedding for meta-learning." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[8] Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." arXiv preprint arXiv:1708.00489 (2017).