Test Accuracy vs. Generalization Gap: Model Selection in NLP without Accessing Training or Testing Data

Yaoqing Yang
Dartmouth College

Ryan Theisen
University of California, Berkeley

Liam Hodgkinson
University of Melbourne

Joseph E. Gonzalez
University of California, Berkeley

Kannan Ramchandran
University of California, Berkeley

Charles H. Martin
Calculation Consulting

Michael W. Mahoney
University of California, Berkeley

ABSTRACT

Selecting suitable architecture parameters and training hyperparameters is essential for enhancing machine learning (ML) model performance. Several recent empirical studies conduct large-scale correlational analysis on neural networks (NNs) to search for effective *generalization metrics* that can guide this type of model selection. Effective metrics are typically expected to correlate strongly with test performance. In this paper, we expand on prior analyses by examining generalization-metric-based model selection with the following objectives: (i) focusing on natural language processing (NLP) tasks, as prior work primarily concentrates on computer vision (CV) tasks; (ii) considering metrics that directly predict *test error* instead of the *generalization gap*; (iii) exploring metrics that do not need access to data to compute. From these objectives, we are able to provide the first model selection results on large pretrained Transformers from Huggingface using generalization metrics. Our analyses consider (I) hundreds of Transformers trained in different settings, in which we systematically vary the amount of data, the model size and the optimization hyperparameters, (II) a total of 51 pretrained Transformers from eight families of Huggingface NLP models, including GPT2, BERT, etc., and (III) a total of 28 existing and novel generalization metrics. Despite their niche status, we find that metrics derived from the heavy-tail (HT) perspective are particularly useful in NLP tasks, exhibiting stronger correlations than other, more popular metrics. To further examine these metrics, we extend prior formulations relying on power law (PL) spectral distributions to exponential (EXP) and exponentially-truncated power law (E-TPL) families.[1]

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

[1]This is the conference version of a paper that appeared in technical report version as "Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data" [48]; the title is different due to the conference submission policy.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599518

KEYWORDS

Model selection, model quality prediction, Transformers, weight matrix analytics, generalization metrics

ACM Reference Format:
Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. 2023. Test Accuracy vs. Generalization Gap: Model Selection in NLP without Accessing Training or Testing Data . In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599518

1 INTRODUCTION

Selecting the optimal hyperparameters, such as those for training or model size, is a critical phase in the ML pipeline. Motivated by the importance of model selection, recent years have seen a wide array of large-scale empirical studies on the various metrics used to predict the test-time performance of ML models [9, 17, 26, 27]. These *generalization metrics* have been applied in a wide variety of data science tasks, including predicting the quality of pretrained learning models [23, 27], designing effective training procedures [11, 14], improving network efficiency [5, 8], quantifying model robustness [41, 47], improving ensemble learning techniques [12, 13], analyzing and improving large-scale machine learning contests [26], and so on. They are typically studied using correlational analysis, measuring how strongly each metric correlates with (and therefore, can predict) model performance. In this regard, several recent works point out the deficiencies of existing generalization metrics, including a lack of "robustness" to the changes of environmental hyperparameters [9, 17] (such as data, neural network architecture and training schemes), or the *Simpson's paradox* that generalization metrics perform differently (i.e., predict opposite trends) when applied to each sub-part of a collection of learning models or to the holistic study [26]. Another drawback is the over-reliance on CV models, which are relatively well-explored, and are not always representative of other types of tasks. With few exceptions [27, 31, 46], systematic studies in other fields, such as NLP, are largely missing.

Generalization metrics for model selection in NLP. The objective of this work is to provide a systematic study of generalization metrics in NLP, addressing several deficiencies in prior studies [9, 17, 27]. Compared to CV, model selection in NLP has several

important differences that require careful consideration. For example, the training data from standard CV benchmarks can often be easily obtained, while large language model datasets are typically web-scale and are challenging to access. Therefore, generalization metrics that can assess the quality of learning models *without access to data* are ideal for NLP. In this paper, we focus on generalization metrics that do *not* need access to data, which is useful for evaluating pretrained NLP models [45]. Indeed, recent work has demonstrated that access to training or testing data should not be necessary for assessing the model quality of learning models [27], though these findings have yet to be evaluated at scale in the NLP domain. Furthermore, it is typically infeasible to train NLP models to interpolate the (frequently large) training set. Contrary to common practice for CV models, the training error on NLP datasets is often much larger than zero. This becomes an issue when applying most existing generalization metrics as they compare models through the *generalization gap* (i.e., the difference between training and test performance) rather than the test error itself. Metrics that focus on ranking the generalization gap include most of the well-known metrics in CV, such as those based on the PAC-Bayesian framework [28, 33] and margins [3, 16, 37].

To illustrate the issue, consider selecting between two models with test errors e_1, e_2, training errors l_1, l_2, and generalization gaps $g_1 = e_1 - l_1$ and $g_2 = e_2 - l_2$. Assuming a generalization metric can *rank* the generalization gap perfectly (which is often the focus of prior studies on generalization metrics [9, 15, 17]) [2], we know only that one model has a larger training-test gap than another ($g_1 > g_2$). For these two models, even if we have access to both models' exact training errors l_1, l_2, we still cannot determine which model exhibits smaller test error: if $l_1 < l_2$, we cannot determine whether $l_1+g_1 > l_2+g_2$ unless we know the training-test gaps g_1, g_2 *explicitly*. Therefore, if our objective is to construct a metric that correctly predicts model performance, rank correlation with the generalization gap is insufficient. In this paper, we aim to study how generalization metrics rank correlate with model quality, for which we use test error as a close approximation. As we will demonstrate (in Figure 4), rank correlation with the generalization gap indeed does not imply rank correlation with model quality in practice, and in fact often orders models in the opposite order of their test errors. From a practical point of view, for NLP tasks, we prefer generalization metrics that can directly predict trends in test error (or similar evaluation metrics in NLP, such as the test BLEU score [36]) rather than trends in the generalization gap.

Naturally, we cannot expect a metric to be universally correlated with test error if evaluating the metric does not need data. However, within certain classes of models (e.g., stages of training in one model or across pre-trained models), they may be effective at diagnosing model quality. With these objectives in mind, among the generalization metrics in the literature, we take particular interest in those derived from the heavy-tail self regularization (HT-SR) theory [23, 25] due to reasons summarized in the following:

[2]As the report of the NeurIPS 2020 Competition on Predicting Generalization in Deep Learning [15] points out, the generalization metric "should" be able to order models' performance in a way similar to the generalization gap, and thus one hopes that it can be used for model selections or neural architecture search. However, see Martin and Mahoney [26] for a detailed exposition of issues and problems with this.

> We choose HT-SR generalization metrics for model selection in NLP because they (i) predict test error directly instead of the generalization gap and (ii) do not require access to training (or testing) data.

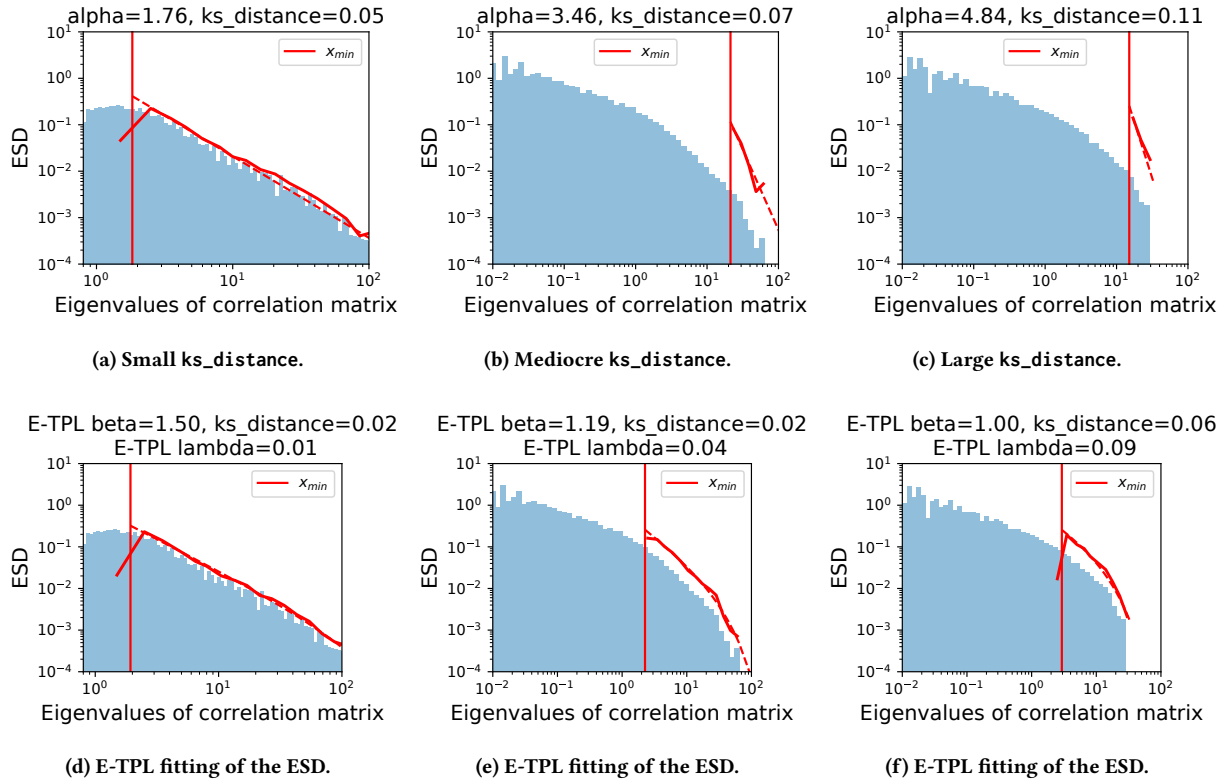
HT-SR theory and shape metrics. The core principle of HT-SR theory is that HT structures arise naturally in the ESDs of the weight matrices [3] as the result of extracting various correlations in data during optimization [23–27]. Its primary practical consequence is that by estimating the PL coefficient from the ESDs (requiring only weights), one can predict model quality, as smaller coefficients are reported to correspond to higher test accuracy. However, these estimators can be unstable, and so one must be careful not to rely on them alone. The quality of the PL fit itself should also point to similar conclusions [25], which can be a sanity check.

The principles of HT-SR theory extend beyond fitting the PL coefficient, however, as ESDs can take many forms. To this end, we study three different types of distributions to fit to the ESDs of weight matrices, including power laws (PL) in Eqn. (1), exponentially truncated power laws (E-TPL) in Eqn. (2), and exponential laws (EXP) in Eqn. (3). These are all commonly considered families of distributions in classical studies of PL [6], and it is often hard in practice to predict which family fits data the best. Figure 1 shows examples of comparing different HT fittings on the same ESD. Following Martin and Mahoney [26], we refer to the various metrics derived from HT-SR as *shape metrics*.

Contributions. The following summarizes our main contributions.

- Deviating from prior work examining generalization metrics in CV [9, 17], we provide the first systematic correlational analysis on various generalization metrics in NLP. Our detailed studies include:
  - considering 360 transformers trained on WMT14 [4] with varying hyperparameters, and eight families of pretrained SOTA transformers downloaded from Huggingface [45], including BERT [18], GPT2 [38], ALBERT (both v1 and v2) [19], etc;
  - providing the first systematic study of applying generalization metrics to the model selection of Transformers without any training/validation/testing data;
  - measuring the correlation between 28 generalization metrics and the model quality (measured by test-time performance) over three different model classes: (i) models trained with the optimal hyperparameters, (ii) a single model at different stages of training, and (iii) a model trained with different hyperparameters (similar to Jiang et al. [17], Martin and Mahoney [26].)
- We revisit prior findings on data-dependent metrics motivated by margins and PAC-Bayesian bounds [9, 17], finding that while these metrics perform well in predicting the *generalization gap*, none of them satisfactorily predicts test error directly.
- When applied appropriately, we find that HT-based shape metrics consistently perform better than scale metrics (or norm-based metrics) for predicting model quality.

[3]The ESD of a weight matrix $\mathbf{W}$ refers to the empirical density of the eigenvalues of the squared weight matrix $\mathbf{W}^\top\mathbf{W}$. See "Preliminary of ESDs of weight matrices" at the end of the Introduction.



(a) Small `ks_distance`. **(b) Mediocre `ks_distance`.** **(c) Large `ks_distance`.**

(d) E-TPL fitting of the ESD. **(e) E-TPL fitting of the ESD.** **(f) E-TPL fitting of the ESD.**

Figure 1: Comparing PL and E-TPL fitting. (First row). Good, mediocre, and bad PL fittings measured by the `ks_distance`. (Second row). E-TPL fitting of the ESD on the same column. Blue histograms represent the ESDs. Solid vertical lines represent the lower threshold $x_{\min}$ of the PL distribution found by the fitting procedure. Solid curves represent ESDs truncated using $x_{\min}$, and dashed curves represent the fitted HT distributions.

- We extend prior studies on HT-SR theory and investigate alternative models to fit heavy-tail/light-tail distributions. Our results show that E-TPL fits are comparatively robust alternatives to PL fits on suboptimally-trained models.

A more detailed empirical evaluation may be found in the arXiv version [48], including corroborating results on Wikitext-103, Reddit and MNLI, definitions of all the generalization metrics used in this paper, and results comparing various ways of measuring rank correlations, such as Spearman's rank correlation and Kendall's tau. In order that our results can be reproduced and extended, we have open-sourced our code.[4]

Preliminary of ESDs of weight matrices. Consider a NN with d layers and corresponding weight matrices $\mathbf{W}_1$, $\mathbf{W}_2$,..., $\mathbf{W}_d$. For each weight matrix $\mathbf{W}_i$ with shape $N \times M$, assume without loss of generality that $N \geq M$ (otherwise, consider $\mathbf{W}_i^\top$). We define the correlation matrix as $\mathbf{X}_i = \mathbf{W}_i^\top \mathbf{W}_i$, and denote the eigenvalues of $\mathbf{X}_i$ as $\{\lambda_j\}_{j=1}^M$, so that $\lambda_j = \sigma_j^2$, where $\{\sigma_j\}_{j=1}^M$ are the singular values of $\mathbf{W}_i$. Furthermore, we use $\lambda_{i,\max}$ to denote the maximum eigenvalue of the correlation matrix $\mathbf{X}_i$. The ESD (empirical spectral density) of the weight matrix $\mathbf{W}_i$ refers to the empirical density of the eigenvalues of $\mathbf{X}_i$, typically represented through a histogram.

We let $p(x)$ denote the density function to fit the ESD taking values in the interval $(x_{\min}, x_{\max})$. For a power law, p satisfies

$$p(x) \propto x^{-\alpha}, \quad x_{\min} < x < x_{\max}. \tag{1}$$

From Martin and Mahoney [26], $x_{\max}$ is chosen to be the maximum eigenvalue of the empirical correlation matrix. However, $x_{\min}$ is a variable to be optimized to improve the quality of PL fitting, and it is not equal to the minimum eigenvalue in general.

2 HEAVY-TAIL SELF-REGULARIZATION

Here, we provide a brief overview of the HT-SR theory, and discuss several metrics that can be derived from it. According to HT-SR theory, the ESDs of the weight matrices become more heavy-tailed during training as they become increasingly correlated. One can quantify the extent of these correlations by fitting a PL to the ESD of a weight matrix, for example, by using the open-source `WeightWatcher` tool [5][27]. After computing the ESD of a weight matrix, we use the maximum likelihood estimate from Alstott et al. [1] to fit the PL distribution, the specific form of which has been defined in (1). Let `PL_alpha` denote the PL coefficient averaged over layers; effectively the slope of the tail of the ESD of the pooled weights, on a log-log scale.

[4]https://github.com/nsfzyzz/Generalization_metrics_for_NLP

[5]https://github.com/CalculatedContent/WeightWatcher

Correctly identifying and fitting PL distributions is well-known to be a challenge in practice. For example, a density that appears as a straight line on a log-log scale plot need not follow a power law, as there are many other distributions that could show a similar behavior, including lognormal and exponential-type distributions [6]. Nested distributions such as E-TPL, which combine the pure PL and other distributional assumptions, can often improve the quality of fitting [1, 6]. Therefore, in addition to PL (defined in (1)), we consider several other distribution classes from the literature.

- (`E_TPL_lambda` and `E_TPL_beta`) The ESDs are assumed to take a "nested" form in the interval $(x_{\min}, x_{\max})$.

$$p(x) \propto x^{-\beta} \exp(-\lambda x),\ x_{\min} < x < x_{\max}. \tag{2}$$

  After fitting the E-TPL, we call the exponential truncation coefficient λ the `E_TPL_lambda` metric, and we call the PL coefficient the `E_TPL_beta` metric.
- (`EXP_lambda`). The ESDs are assumed to take the following form, in the interval $(x_{\min}, x_{\max})$.

$$p(x) \propto \exp(-\lambda x),\ x_{\min} < x < x_{\max}. \tag{3}$$

  After fitting the EXP, we call the exponential coefficient λ the `EXP_lambda` metric.

For more details of the various metrics considered in this paper, see Table 1. All of the metrics derived from HT-SR do *not* require access to data, and they are relatively cheap to compute. Our primary comparisons are between shape metrics (derived from HT-SR), and scale metrics (mostly norm-based). Scale metrics are mostly studied in prior work [9, 17], while shape metrics have received less attention. For the precise definitions of these metrics, see Appendix A of our full report online [48].

Issues of PL fitting. It is well-known that subtle issues can arise when fitting the ESDs [1, 6, 22, 26]. To best mitigate these issues in PL fits, we adopt the fitting strategies used in `WeightWatcher` [22]. For example, as in Clauset et al. [6], it is common to choose the lower threshold $x_{\min}$ which coincides with the best quality fit under the Kolmogorov–Smirnoff statistic defined as:

$$\mu_{\texttt{ks\_distance}} = \sup_x |F^*(x) - S(x)|, \tag{4}$$

where $F^*(x)$ is the distribution of the estimated PL fit to the ESD of the weight matrix, and $S(x)$ is the ESD itself. We will refer to (4) as `PL_ks_distance`, or `E_TPL_ks_distance` when the fitting is E-TPL. However, this method is time-consuming, especially for E-TPL as there are two parameters to fit. Instead, we adopt the *fix-finger method* (see `WeightWatcher`) which selects $x_{\min}$ as the peak of the ESD when fitting E-TPLs. More than a simple speed improvement, we find this method also yields more stable results.

Comparing PL and E-TPL fitting. Referring to Figure 1, we now discuss how E-TPL could partially address these fitting issues. On the first row of Figure 1, we show three typical cases of PL fitting. In Figure 1a, the log-log scale reveals a "linear region" of the histogram, which the PL fitting correctly locates. The quality of fit, measured by the `ks_distance`, is within a typical range, as reported in Table 6 of Martin and Mahoney [25]. In Figure 1b and Figure 1c, the ESDs do not exhibit a clear linear region on the log-log scale. Following Martin and Mahoney [25], it is ill-advised to consider metrics derived from a PL fit in these scenarios. In practice, this typically occurs when `PL_alpha` > 4 (e.g., see Figure 1c). On the other hand, in these two cases, the corresponding E-TPL fits (shown on the second row in Figure 1) still closely match the empirical density function (see Figure 1e and Figure 1f), and the `ks_distance` on the second row using a E-TPL fit is smaller than that for the PL fit on the first row, even when the fit on the second row clearly covers a larger part of the ESD. In these two cases, the `E_TPL_lambda` plays a similar role as the `PL_alpha` in PL fitting, and provides an effective alternative when the ESD does not exhibit a proper PL.

3 EMPIRICAL RESULTS

3.1 Experimental setup

Dataset. In Section 3.2, we study models trained on the WMT14 German to English (DE-EN) dataset [4], commonly used as a benchmark for neural machine translation [10, 35, 40, 43]. WMT14 consists of 4.5 million sentence pairs for training.

Hyperparameters. To conduct correlational analysis, and to capture the relationship between the generalization metrics and model quality in different settings, we vary several hyperparameters: the number of samples (either 160K, 320K, 640K, 1.28M, 2.56M samples), the initial learning rate during training (across eight different rates), the model width (embedding dimension either 256, 384, 512, 768, or 1024), and the model depth ({4, 5, 6, 7, 8}-layer transformers). Similar to prior works on correlational analysis [17] for model selection, we construct a high-dimensional grid of different hyperparameters $\mathbf{\Theta} = \{(\theta_1, \ldots, \theta_K) : \theta_1 \in \mathbf{\Theta}_1, \ldots, \theta_K \in \mathbf{\Theta}_K\}$, so that we can compare models when one of the hyperparameters is varied. Two separate high-dimensional grids with dimension $K = 3$ are considered: (1) sample×learning rate×width; (2) sample×learning rate×depth. Each grid contains 5×8×5=200 of these training settings. In total, there are 360 trained models because the two high-dimensional grids overlap each other, and 40 models belong to both grids. We will conduct three correlational analyses in the following to evaluate model selection performance.

Task one, correlation evaluated on optimally trained models. In the first task (Section 3.2.1), we measure the (rank) correlation between model quality and generalization metrics on models trained with the optimal choice of training hyperparameters, that is, if we are allowed to grid-search the best training hyperparameters, can we predict the best data size or model size parameters?

Task two, correlation in time. In the second task (Section 3.2.2), we track BLEU score and generalization metrics during training, assessing time-wise correlation to model quality. This task has been considered in the literature [3], and from a practical point of view, capturing the time-wise dependence during training could potentially lead to better early stopping and regularization methods.

Task three, correlation when a single hyperparameter is varied. In the third task (Section 3.2.3), we study the relationship between the model quality and the generalization metrics when a single hyperparameter is varied. Metrics that achieve a high (rank) correlation for all the hyperparameters are good candidates for model selection.

Training and model setup. For the details of training Transformers on WMT14, see Appendix B of the online report [48].

Name	Ref	Need initial weights?	Scale or shape	Need data?	Need gpu?	Predicting model quality or generalization gap?
param_norm	[17]	No	Scale	No	No	Generalization gap
fro_dist	[17]	Yes	Scale	No	No	Generalization gap
log_norm	[25]	No	Scale	No	No	Generalization gap
log_spectral_norm	[26]	No	Scale	No	No	Generalization gap
dist_spec_int	[17]	Yes	Scale	No	No	Generalization gap
path_norm	[34]	No	Scale	No	No	Generalization gap
mp_softrank	[25]	No	Scale/Shape	No	No	Model quality
stable_rank	[25]	No	Scale/Shape	No	No	Model quality
PL_alpha	[25]	No	Shape	No	No	Model quality
E_TPL_beta	This paper WeightWatcher	No	Shape	No	No	Model quality
E_TPL_lambda	This paper WeightWatcher	No	Shape	No	No	Model quality
EXP_lambda	This paper WeightWatcher	No	Shape	No	No	Model quality
PL_ks_distance	[25]	No	Shape	No	No	Model quality
E_TPL_ks_distance	This paper [25]	No	Shape	No	No	Model quality
alpha_weighted	[25]	No	Hybrid	No	No	Model quality
log_alpha_norm	[26]	No	Hybrid	No	No	Model quality
inverse_margin	[17]	No	Scale	Yes	Maybe	Generalization gap
log_prod_of_spec_over_margin	[3, 37]	No	Scale	Yes	Maybe	Generalization gap
log_sum_of_spec_over_margin	[3, 37]	No	Scale	Yes	Maybe	Generalization gap
log_prod_of_fro_over_margin	[3, 37]	No	Scale	Yes	Maybe	Generalization gap
log_sum_of_fro_over_margin	[3, 37]	No	Scale	Yes	Maybe	Generalization gap
path_norm_over_margin	[34]	No	Scale	Yes	Maybe	Generalization gap
pacbayes_init	[32]	Yes	Scale	Yes	Yes	Generalization gap
pacbayes_orig	[32]	No	Scale	Yes	Yes	Generalization gap
pacbayes_flatness	[32]	No	Scale	Yes	Yes	Generalization gap
pacbayes_mag_init	[17]	Yes	Scale	Yes	Yes	Generalization gap
pacbayes_mag_orig	[17]	No	Scale	Yes	Yes	Generalization gap
pacbayes_mag_flatness	[17]	No	Scale	Yes	Yes	Generalization gap

Table 1: Overview of the generalization metrics considered in this paper. We focus on the *shape* metrics derived from the ESDs of weight matrices. Due to the space constraint, detailed definitions of these metrics are presented in Appendix A of our full version online [48].

3.2 Correlational analyses on Transformers trained in different settings

In this subsection, we study 28 generalization metrics (with details provided in Table 1) and examine their correlations with BLEU score [36], the most commonly used metric to evaluate machine translation [6]. Note that BLEU score here is used as a close approximation of model quality, mimicking the role of test accuracy in image classification. We also consider correlation between these metrics and the generalization gap, defined as the BLEU score for training data subtracted by the BLEU score for test data. We intend to find generalization metrics that strongly correlate with model quality instead of the generalization gap.

3.2.1 Task one: Evaluating correlations on optimally trained models only. Here, we group models using the number of training samples, and select the best model from each group when the model depth and the learning rate are varied. In Figure 2, each curve represents a group of models trained with a certain number of training samples. The black star on each curve represents training with optimal hyperparameters (learning rate and depth in our setting), obtained by searching for the optimum on a third-order polynomial fit of each curve. From Figure 2, we see that the shape metrics correctly predict the model quality for models trained with the optimal training hyperparameters, i.e., the BLEU scores should be higher when the metric values are smaller on the optimal models represented using black stars. Since all six shape metrics show similar trends, a pairing of these metrics can be considered as a sanity check.

Comparison with scale metrics. We compare scale metrics and shape metrics in Section 3.2.3 in our full report [48]). We show that shape metrics predict the correct trends in test BLUE scores, while scale metrics predict wrongly because they are correlated with the generalization gap.

Remark 3.1. Figure 2 points out an important but subtle issue in empirically evaluating the HT-SR theory. In Figure 2, one can make a model less well-trained—and artificially anti-correlate the

[6] Several empirical metrics have been designed to measure the quality of text generation, such as BERTScore [50] and BARTScore [49]. Our work is different because we do not need any data, and we do model selection using the ESDs of weight matrices only. BERTScore and BARTScore evaluate the text quality, and thus they need source or reference texts generated by humans. These metrics can serve as alternatives to BLEU, which is viewed as ground truth in our work.

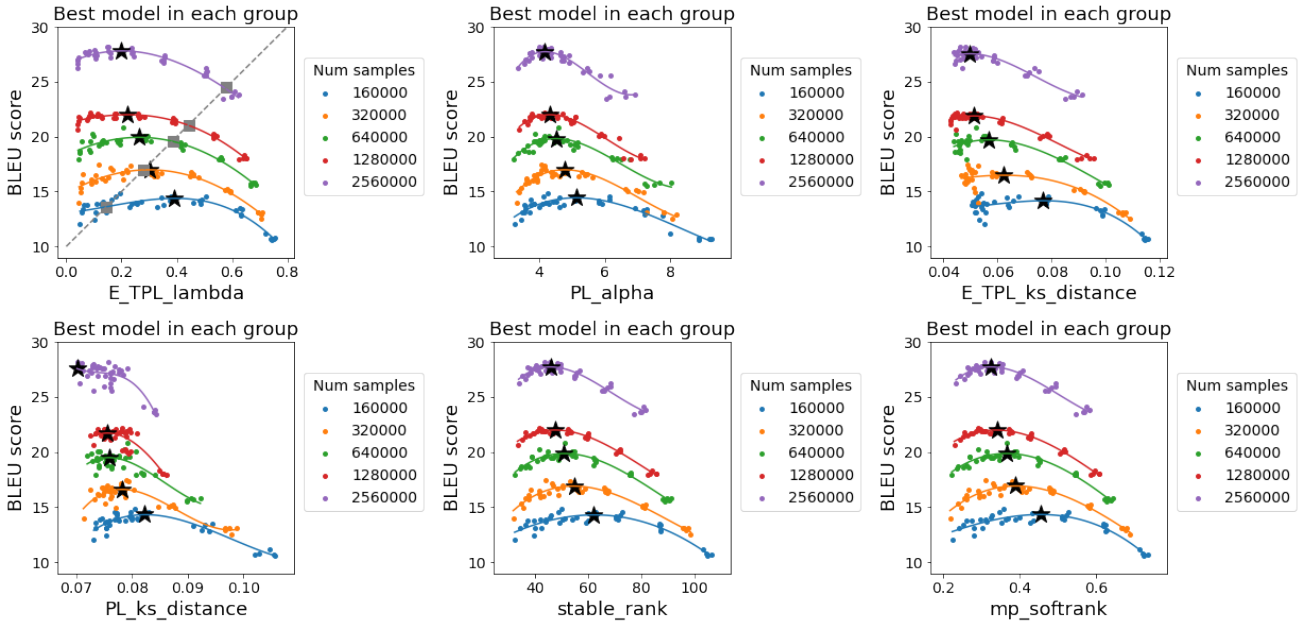


Figure 2: BLEU-score vs. six shape metrics for 200 Transformers trained on WMT14 with varying hyperparameters. HT-SR theory applies for optimally-tuned models (black stars), that is, for optimally-tuned models indicated by the black stars, models that have better BLEU scores exhibit heavier-tailed ESDs. For suboptimal models, the HT-SR metrics can be anti-correlated with model quality, see e.g. the grey dotted line in the first subfigure.

generalization metric with the task accuracy. For example, see the gray dotted line in the first subfigure in Figure 2.

3.2.2 Task two: Time-wise correlations and rank correlation results. In this subsection, we study time-wise correlation between our chosen metrics and the BLEU scores.

`E_TPL_lambda` tracks the BLEU score. As a warm-up, we consider how well the `E_TPL_lambda` metric defined in (2) tracks the BLEU score (recalling that `E_TPL_lambda` assumes the ESDs follow E-TPLs). We use training with and without dropout to study the effect of training schemes, and we consider different quantities of data to test robustness when the size of data changes. In Figure 3, the first row considers models trained with dropout, while the second row considers models trained without dropout. The multiple columns track `E_TPL_lambda` and the BLEU score throughout training for different amounts of data. We can see that `E_TPL_lambda` not only successfully tracks BLEU scores but also differentiates underfitting (first row, with dropout) from overfitting (second row, without dropout) in this experiment.

Shape metrics predict model quality, while scale metrics predict the generalization gap. Now we consider the rank correlations between our chosen metrics and the test BLEU score. The rank correlations are evaluated across training, i.e., for each of the 360 settings of the hyperparameters, we calculate the Spearman's rank correlation between BLEU scores and the values of each generalization metric over all epochs. The summarized results are presented in Figure 4a. A positive Spearman's rank correlation (with BLEU) suggests that the generalization metric is useful in tracking BLEU during training. A negative Spearman's rank correlation, on the other hand, implies that the metric often gives the incorrect prediction. In Figure 4a, we use the average rank correlations for all settings to study the effectiveness of each metric, and present 25% quantile rank correlations to indicate robustness across runs.

In Figure 4a, we find shape metrics, such as `E_TPL_ks_distance`, `EXP_lambda`, `E_TPL_lambda`, and `E_TPL_beta`, exhibit some of the highest rank correlations with BLEU score. The `EXP_lambda` metric, which assumes a EXP distribution on the ESDs, achieves the highest median rank correlation, while the `E_TPL_lambda` metric, which assumes a E-TPL distribution on the ESDs, achieves the second highest.

In Figure 4b, we plot the rank correlations to the generalization gap across our chosen metrics. While it is encouraging that most existing generalization metrics yield correct predictions, as previously discussed, correct predictions of the generalization gap do *not* imply accurate predictions on the best-performing models here.

Details of the rank correlation calculations. When calculating the rank correlation with the test accuracy, we associate a negative sign to all the generalization metrics, i.e., a positive rank correlation in Figure 4a means that a generalization metric is negatively correlated with the BLEU score. We use this procedure to follow the conventional wisdom that a smaller value of the complexity metric leads to better generalization [17]. On the other hand, for Figure 4b, a positive rank correlation means that the metric is positively correlated with the generalization gap. Thus, for both Figure 4a and 4b, a strong positive correlation corresponds to the expected trend.

Can we utilize anti-correlation for prediction? One may ask if the anti-correlation shown in Figure 4b implies that scale metrics can also predict model quality. Indeed, from Figure 4b alone, it

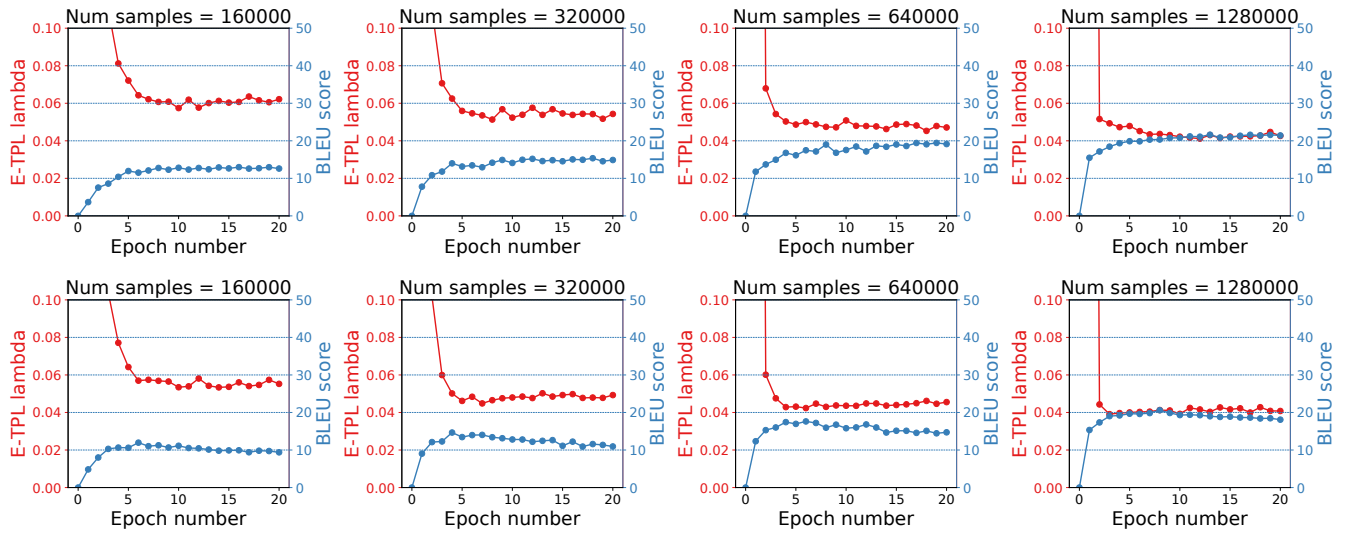
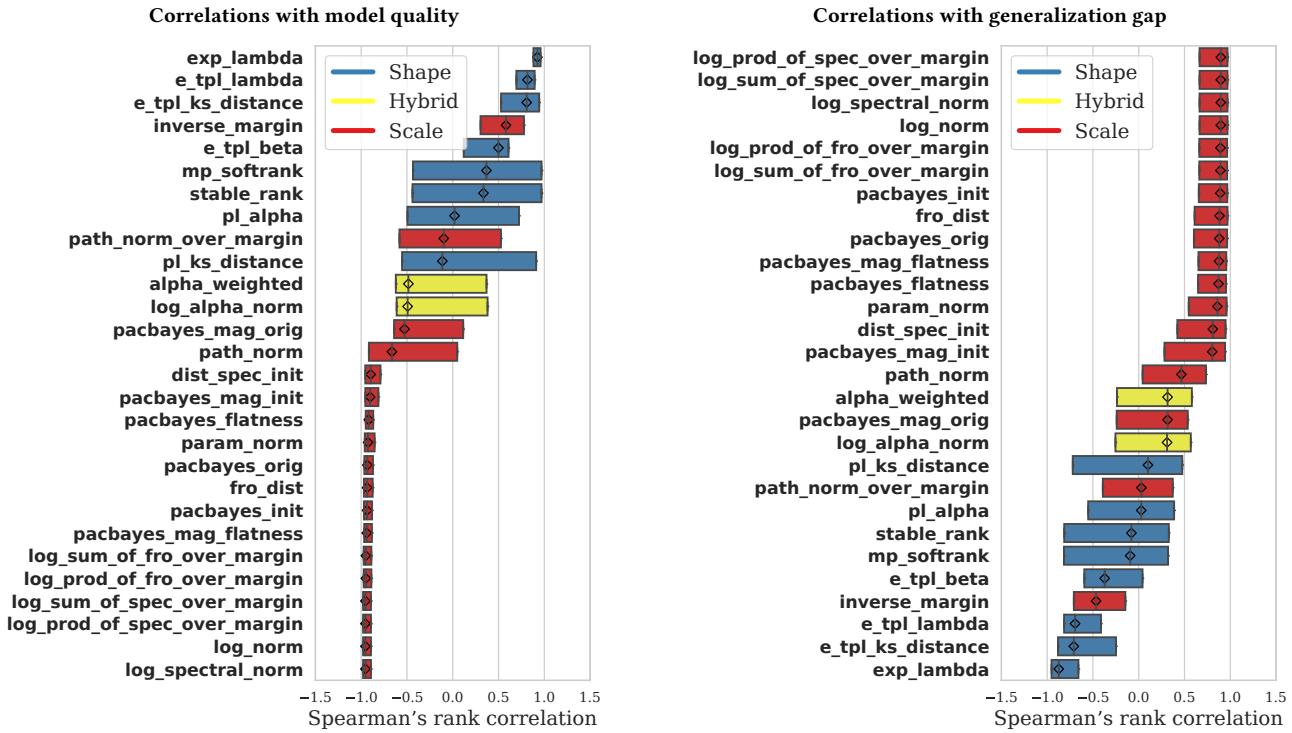


Figure 3: `E_TPL_lambda` closely tracks the BLEU score, i.e., BLEU score increases when the `E_TPL_lambda` drops. Results are shown for Transformers trained on WMT14 with different number of samples. (First row). Training with dropout 0.1. (Second row). Training without dropout.



(a) Correlations with model quality. Spearman's rank correlation between various generalization metrics and BLEU.

(b) Correlations with generalization gap. Spearman's rank correlation between various generalization metrics and generalization gap.

Figure 4: Comparing multiple generalization metrics for predicting BLEU score (on the left) or the generalization gap (on the right). Lines on each box delineate the 25/50/75 percentiles of the rank correlations in 360 different settings (including different amount of data, different network depths, different network widths, and different initial learning rates).

seems that one can negate the predicted results of scale metrics to obtain an accurate prediction. However, note this strong negative correlation of scale metrics only holds in this one particular scenario. In other scenarios, such as in Dziugaite et al. [9], Jiang et al. [17], the correlation is strong in the other direction. Broadly speaking, if a particular theory says that a quantity should go up with model quality, and it goes down sometimes instead, then the theory is incomplete, regardless of how strong the correlation is. A prominent claim in our paper is that the correlation between test error and the generalization gap can sometimes be reversed. Therefore, it is insufficient to study metrics that have a large rank correlation with the generalization gap.

3.2.3 Task three: evaluating correlation when a single hyperparameter is varied. In this subsection, we assess whether the generalization metrics can predict trends in BLEU score when a single hyperparameter is changed. Specifically, for a hyperparameter space $\mathbf{\Theta} = \{(\theta_1, \ldots, \theta_K) : \theta_1 \in \mathbf{\Theta}_1, \ldots, \theta_K \in \mathbf{\Theta}_K\}$, we consider each one-dimensional slice of the form

$$\{(\theta_1, \ldots, \theta_K) : \theta_i \in \mathbf{\Theta}_i \text{ while other parameters } \theta_j, j \neq i \text{ are fixed}\},$$

and we calculate the rank correlation using the models in each such slice. Then, we aggregate the rank correlations from all the one-dimensional slices and plot the distributions of the rank correlations. For example, if we evaluate the trends when the initial learning rate is varied, we choose $\mathbf{\Theta}_i$ to be the set of eight different initial learning rates mentioned in Section 3.1, "Hyperparameters". As another example, we can define $\mathbf{\Theta}_i$ to be the set of five different numbers of samples to study the (rank) correlation when the number of samples is varied.

Similar to Figure 4, we provide the rank correlation results on both the test BLEU scores and the generalization gap. See Section 3.2.3 of our report online [48]. Again, shape metrics have better rank correlations with model quality, while scale metrics are better correlated with the generalization gap.

Corroborating results. We extend our empirical evaluations to other datasets and evaluation methods. First, we consider pretrained Huggingface Transformers in Section 3.3, providing model selection results in a broad range of NLP tasks. Then, we consider three other language processing tasks trained with different Transformers, including

- Roberta [21] trained on the masked language modeling task using Wikitext-103 [29], and then finetuned on MNLI [44];
- Six-layer base Transformers trained on the language modeling task using the Wikitext-103 dataset [29];
- Six-layer base Transformers trained on the next-word prediction task using the Reddit dataset, following the implementation in Bagdasaryan et al. [2].

All extended results can be found in our online report [48]. Also, in [48], we provide additional results on conducting correlational analysis using Kendall's tau instead of Spearman's rank correlation.

Computational cost and carbon emission. We believe it is extremely important that papers relying on large-scale empirical analysis accurately report the computational cost. The overall training cost is 7301.66 GPU hours. We use GPU nodes with TITAN RTX for our training. The overall carbon emission depends on carbon efficiency. Using the default values from the online Machine Learning Emissions Calculator[7], the total emissions are estimated to be 883.21 kg CO2 eq.

Model series	Models
BERT [18]	BERT {Tiny, Mini, Small, Base, Large}
Smaller BERT [42]	24 smaller BERT models (English, uncased, trained with WordPiece masking)
GPT2 [38]	GPT2 {Original, Medium, Large, XL}
ALBERTv1 [19]	ALBERT-v1 {base, large, xlarge, xxlarge}
ALBERTv2 [19]	ALBERT-v2 {base, large, xlarge, xxlarge}
T5 [39]	T5 {small, base, large}
DialoGPT [51]	DialoGPT {small, medium, large}
FlauBERT [20]	FlauBERT {small, base, large}
Funnel Transformer [7]	FunnelModel {small, medium, intermediate, large, xlarge}

Table 2: Pretrained Transformers considered in this paper.

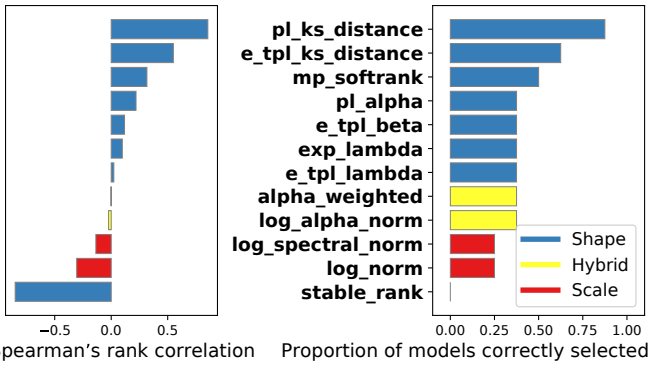
3.3 Selecting Huggingface Transformers

Finally, we evaluate generalization metrics on the model selection task of pretrained Transformers. This section presents the first systematic study of applying generalization metrics to the model selection of Transformers without any training/validation/testing data. In our study, eight series of models downloaded from Huggingface [45] are considered—see Table 2. We also include 24 BERT models from the "Smaller BERT" series [42] produced from a "pretrained distillation" pipeline that combines masked language modeling pretraining [18] and knowledge distillation from a single BERT teacher model. In total, there are 51 pretrained Transformers.

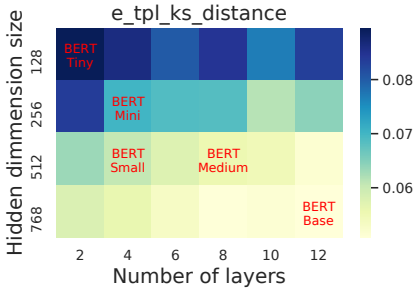
We report rank-correlations averaged over these 8 model series in Figure 5a (left subplot), i.e., larger/deeper models should have smaller generalization metric values. Again, we find that the shape metrics outperform scale metrics (except for `stable_rank`, which is strongly influenced by the size of the weight matrix). The hybrid models achieve performance in-between the shape and scale metrics. In Figure 5a (right subplot), we compare different metrics in their ability to select the *best model*. That is, we report for each metric the proportion that the best model is selected from one model series when this metric is used as the model selection criterion. Note that the rankings of metrics on the two subplots in Figure 5a are the same.

From Figure 5a, we can see that, while the shape metrics perform better than scale metrics, none show a particularly strong rank correlation. To understand this, we examine the "Smaller BERT" series [42], which contains a more fine-grained structure of different model sizes. Specifically, these models are arranged in a 4-by-6 grid, where 6 represents {2,4,6,8,10,12} transformer layers and 4 means different hidden embedding sizes {128,256,512,768}. From Figure 5b, we see that the `E_TPL_ks_distance` correctly predicts the trend that wider and deeper models perform better. On the other hand, from Figure 5c, `E_TPL_lambda` correctly predicts that wider models are better, but incorrectly predicts that shallower models are better (yet another form of Simpson's paradox in a data set of neural network model quality; see also Martin and Mahoney [26]).

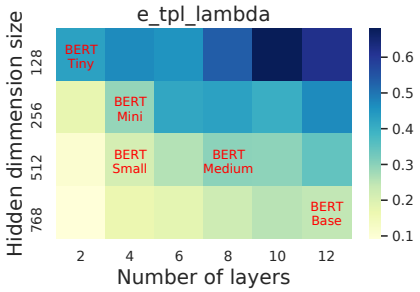
[7] https://mlco2.github.io/impact/#compute



(a) Model selection on Huggingface Transformers. Metrics on the left and right are aligned.



(b) `E_TPL_ks_distance` evaluated on BERT models of different size.



(c) `E_TPL_lambda` evaluated on BERT models of different size.

Figure 5: Generalization metrics evaluated on pretrained Transformers. (a) Model selection results on eight Huggingface Transformer model series: BERT, GPT2, ALBERTv1, ALBERTv2, T5, DialoGPT, FlauBERT, Funnel Transformer. Left shows the rank correlation averaged over different Transformers. Right shows the proportion of the best Transformers correctly selected using different metrics. Shape metrics outperform scale metric only except `stable_rank` which is strongly affected by the matrix size. (b and c) Evaluating two metrics on the "Smaller BERT" series. While `E_TPL_ks_distance` predicts the correct trends, `E_TPL_lambda` shows the reversed trends with depth.

Another curious observation from Figure 5a is that, for the pretrained transformers, PL metrics, such as `PL_ks_distance` and `PL_alpha`, outperform E-TPL metrics, such as `E_TPL_ks_distance`, `E_TPL_lambda`, and `E_TPL_beta`. This phenomenon may seem surprising as one may expect E-TPL fits to be more flexible than PL fits. These pretrained models are likely trained with much larger datasets and over many more epochs than the models we have otherwise considered. Here, PLs appear to provide a more natural fit. This is further evidence that HT-SR theory is particularly well-suited for evaluating the quality of relatively high-quality models.

4 CONCLUSION

Poor correlations between existing generalization metrics and test-time performance have been reported in prior work [9, 17, 30]. Rather than providing a "lump sum" to rank existing and novel generalization metrics (Figure 4), we evaluated these metrics in several ways: quantifying correlations only on optimally-trained models (Figure 2); examining the time-wise correlation during training (Figure 3); differentiating between the correlation with test accuracy versus generalization gap (Figure 4); providing the first result on model selection of pretrained Transformers using these metrics (Figure 5); and thoroughly investigating the rich correlational structures when different hyperparameters are varied (see the full paper [48]). Our large-scale empirical analyses suggest that popular generalization metrics still exhibit excellent correlations with generalization gap on NLP tasks. However, metrics derived from HT-SR theory appear to be most valuable to large language model practitioners, allowing one to assess pretrained NLP models without requiring training or testing data. Due to their apparent utility and current niche status, we recommend further investigations into these metrics, in particular, to address some of their remaining weaknesses (e.g. for suboptimally-trained models).

Acknowledgements. MM would like to acknowledge the IARPA (contract W911NF20C0035), NSF, and ONR for providing partial support of this work. KR would like to acknowledge support from NSF CIF-1937357, NSF CIF-2007669 and ARO fund 051242-002. JG would like to acknowledge supports from NSF CISE Expeditions Award CCF-1730628, NSF CAREER Award and gifts from Alibaba Group, Amazon Web Services, Ant Group, Ericsson, Facebook, Futurewei, Google, Intel, Microsoft, Nvidia, Scotiabank, Splunk and VMware. WeightWatcher is a publicly-available tool distributed under Apache License 2.0 with copyright held by Calculation Consulting. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

REFERENCES

[1] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 2014. Powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one* 9, 1 (2014), e85777.
[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. 2938–2948.
[3] Peter Bartlett, Dylan Foster, and Matus Telgarsky. 2017. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems* 30 (2017), 6241–6250.

[4] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*. 12–58.

[5] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. 2020. Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspired Perspective. In *International Conference on Learning Representations*.

[6] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[7] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems* 33 (2020), 4271–4282.

[8] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2019. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *IEEE/CVF International Conference on Computer Vision*. 293–302.

[9] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. 2020. In search of robust measures of generalization. *Advances in Neural Information Processing Systems* 33 (2020).

[10] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 489–500.

[11] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

[12] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2019. *Deep ensembles: A loss landscape perspective*. Technical Report Preprint: arXiv:1912.02757.

[13] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Conference on Neural Information Processing Systems*. 8803–8812.

[14] P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. 2018. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*. 876–885.

[15] Yiding Jiang, Pierre Foret, Scott Yak, Daniel M Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. 2020. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976* (2020).

[16] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. 2018. Predicting the Generalization Gap in Deep Networks with Margin Distributions. In *International Conference on Learning Representations*.

[17] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2019. Fantastic Generalization Measures and Where to Find Them. In *International Conference on Learning Representations*.

[18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

[20] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2479–2490.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized BERT pretraining approach*. Technical Report Preprint: arXiv:1907.11692.

[22] Charles H Martin and Michael W Mahoney. 2017. *Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior*. Technical Report Preprint: arXiv:1710.09553.

[23] Charles H Martin and Michael W Mahoney. 2019. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*. 4284–4293.

[24] Charles H Martin and Michael W Mahoney. 2020. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *SIAM International Conference on Data Mining*. SIAM, 505–513.

[25] Charles H Martin and Michael W Mahoney. 2021. Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning. *Journal of Machine Learning Research* 22, 165 (2021), 1–73.

[26] Charles H Martin and Michael W Mahoney. 2021. *Post-mortem on a deep learning contest: a Simpson's paradox and the complementary roles of scale metrics versus shape metrics*. Technical Report Preprint: arXiv:2106.00734.

[27] Charles H Martin, Tongsu Serena Peng, and Michael W Mahoney. 2021. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications* 12, 1 (2021), 1–13.

[28] David A McAllester. 1999. PAC-Bayesian model averaging. In *Annual Conference on Computational Learning Theory*. 164–170.

[29] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. *Pointer sentinel mixture models*. Technical Report Preprint: arXiv:1609.07843.

[30] Vaishnavh Nagarajan and J Zico Kolter. 2019. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems* 32 (2019).

[31] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2019. Deep Double Descent: Where Bigger Models and More Data Hurt. In *International Conference on Learning Representations*.

[32] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. 2017. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems* 30 (2017), 5947–5956.

[33] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. 2018. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *International Conference on Learning Representations*.

[34] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2015. Norm-based capacity control in neural networks. In *Conference on Learning Theory*. PMLR, 1376–1401.

[35] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. 1–9.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[37] Konstantinos Pitas, Mike Davies, and Pierre Vandergheynst. 2017. *Pac-bayesian margin bounds for convolutional neural networks*. Technical Report Preprint: arXiv:1801.00171.

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[40] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. *A simple but tough-to-beat data augmentation approach for natural language understanding and generation*. Technical Report Preprint: arXiv:2009.13818.

[41] Thomas Tanay and Lewis Griffin. 2016. *A boundary tilting persepective on the phenomenon of adversarial examples*. Technical Report Preprint: arXiv:1608.07690.

[42] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Well-read students learn better: On the importance of pre-training compact models*. Technical Report Preprint: arXiv:1908.08962.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[44] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)*. 1112–1122.

[45] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 38–45.

[46] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. 2021. Taxonomizing local versus global structure in neural network loss landscapes. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[47] Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. 2020. Boundary thickness and robustness in learning models. *Advances in Neural Information Processing Systems* 33 (2020).

[48] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E Gonzalez, Kannan Ramchandran, Charles H Martin, and Michael W Mahoney. 2022. *Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data*. Technical Report Preprint: arXiv:2202.02842.

[49] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems* 34 (2021), 27263–27277.

[50] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[51] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.