Generalization Measures for Zero-Shot Cross-Lingual Transfer

Anonymous EMNLP submission

Abstract

Building robust and reliable machine learning systems requires models with the capacity to generalize their knowledge to interpret unseen inputs with different characteristics. Traditional language model evaluation tasks lack informative metrics about model generalization, and their applicability in new settings is often measured using task and language-specific downstream performance, which is lacking in many languages and tasks. To address this gap, we explore a set of efficient and reliable measures that could aid in computing more information related to the generalization capability of language models, particularly in cross-lingual zero-shot settings. Our central hypothesis is that the sharpness of a model's loss landscape, i.e., the representation of loss values over its weight space, can indicate its generalization potential, with a flatter landscape suggesting better generalization. We propose a novel and stable algorithm to reliably compute the sharpness of a model optimum, and demonstrate its correlation with successful cross-lingual transfer. [1]

1 Introduction

Generalization enables models to use prior knowledge to reasonably respond to previously unseen stimuli. Although traditional machine learning evaluation is performed based on a preselected set of prediction or generation tasks, accuracy on many public benchmarks may often not be sufficient to extensively assess the ability to perform well in new settings. Therefore, a majority of researchers have found it worthwhile to investigate measures that could evaluate the generalization capability of models with properties, such as VC dimension (Vapnik and Chervonenkis, 1971), cross-entropy (Shannon, 1948), complexity (Mohri et al., 2012) or variation in parameters during training (Nagarajan and Kolter, 2019). Among these, recent findings support the smoothness in the loss curvature to correlate best with generalization capability (Chaudhari et al., 2019; Petzka et al., 2021; Kaddour et al., 2022), motivating the development of learning methods that induce smoothness in the learning trajectory such that the model becomes more robust; either through data perturbation (Jiang et al., 2020a; Aghajanyan et al., 2021; Liang et al., 2021; Hua et al., 2021; Park et al., 2022; Zheng et al., 2021; Wang et al., 2021; Huang et al., 2021) or by integrating the measure directly to the optimization objective (Izmailov et al., 2018; Jastrzebski et al., 2021; Cha et al., 2021; Foret et al., 2021; Hu et al., 2022; Zaken et al., 2022; Stickland and Murray, 2021). However it might often not be straightforward to compute such measures in high-dimensional feature space in a stable fashion (Nachum et al., 2024).

As models get larger and cover more languages, the possibility of improving the applicability of NLP systems in many under-resourced languages gets more promising. An essential requirement in studying the dynamics of cross-lingual knowledge transfer is to have an evaluation methodology that can reliably measure the model's capability in generalization of knowledge under different scenarios. There is a common hypothesis that states that a model demonstrating an extended flat optimum area of low loss value surrounding the minimized loss is indicative of better generalization capability. In this work, we study the above hypothesis and present the first study to provide methods that can be used for measuring the cross-lingual generalization capability of language models.

- We pick prominent measures that were previously shown to correlate well with generalization performance (Jiang et al., 2020b), such as the Frobenius distance of the learned parameters after training (Nagarajan and Kolter, 2019), the margin between model predictions

[1] Code: https://anonymous.4open.science/r/strikegen-7288

and true labels (Wei et al., 2018) and sharpness in loss minima to test applicability to zero-shot cross-lingual generalization measurement (Keskar et al., 2017; Foret et al., 2021).

- We also extend the formulation of state-of-the-art sharpness computation methods (Keskar et al., 2017; Foret et al., 2021) to provide a sharpness prediction algorithm such that the optimization of the parameters can converge in a more stable fashion.

2 Related work

Loss-landscape Minima One of the most promising indicators of generalization capability to date seems to be related to the form of the loss landscape, in particular, the sharpness in the loss curvature. A potential reason for this fallback is traced to stochastic gradient descent (SGD) (Bottou, 2012) methods which often fall into sharp minima of the loss surface (Keskar et al., 2017; Chaudhari et al., 2019; Wang et al., 2021). Although clear conclusions on the relationship between sharpness and generalization performance, such as whether sharper (Dinh et al., 2017) vs. flatter (Li et al., 2018; Keskar et al., 2017) minima would generally yield better generalization, are still due. The main idea behind these methods is that their objective is to explicitly find flat minima, often using stochastic averaging methods (Polyak and Juditsky, 1992; Izmailov et al., 2018), mini-max or sharpness-aware minimization methods, which can be computed by direct formulation based on the Hessian matrix of the loss function (Chaudhari et al., 2019; Petzka et al., 2021) or Monte-Carlo approximations of the minimizer's neighborhood (Foret et al., 2021; Cha et al., 2021).

Adversarial optimization Comparison of two approaches finds that for NLP tasks, mini-max methods are more competitive over averaging-based optimization (Kaddour et al., 2022). Jastrzebski et al. (2021) hypothesize that regularizing the trace of the Fisher information matrix amplifies the implicit bias of SGD, which prevents memorization. The Fisher information (Fisher, 1925) measures local curvature, so a smaller trace implies a flatter minimum, which gives the model more freedom to reach an optimum. Instead of explicitly minimizing the values of parameters, Foret et al. (2021) propose minimizing both loss and sharpness while optimizing the parameters such that they lie in neighborhoods with low loss values. Perturbation is an auxiliary objective that encourages the model predictions to be similar in the vicinity of the observed training samples (Englesson and Azizpour, 2021), usually by penalizing the KL-divergence between the probability distribution of the perturbed and normal model. Perturbations can be adversarial inputs (Jiang et al., 2020a) or inputs with Gaussian or uniform noise (Aghajanyan et al., 2021). To improve cross-lingual generalization, translations of the input generated by machine translation systems were used as perturbed input (Wang et al., 2021; Zheng et al., 2021). Other work also has found the benefit of enforcing consistency for perturbations within the model in addition to the input distribution (Liang et al., 2021; Hua et al., 2021).

3 Methodology

In this study, we undertake the development of a methodology that could benefit an accurate assessment of the generalization capability of models for the purpose of cross-lingual knowledge transfer into under-resourced languages. This section first presents approaches to improving generalization performance and the selected measures that provide stable results for measuring zero-shot cross-lingual transfer performance.

3.1 Sharpness-based Optimization

We chose the following objective functions as fine-tuning methods for a given pre-trained model as a means of comparison since their main purpose is to enhance the generalization and robustness of models. Following the work of Stickland and Murray (2021), as the two most prominent approaches to mini-max optimization, we include Sharpness-Aware Minimization (SAM) (Foret et al., 2021) and regularization with Fisher Information Matrix (FIM) (Jastrzebski et al., 2021) in our evaluation study on cross-lingual generalization.We also include Multi-view Subword Regularization (MVR) as a perturbation-based optimization method (Wang et al., 2021) which induces stochasticity into the shared subword vocabulary across languages for easing cross-lingual transfer.

SAM (Foret et al., 2021) works on the principle of a mini-max objective function: $\min_w \max_{\|\epsilon\|_2<\rho} L(w+\epsilon)$, which essentially means the optimizing function tries to minimize the maximum loss value in a given radius in loss landscape. Therefore, SAM states that it tries to seek "parameters lying in uniformly low-loss neighborhoods".

Fisher Penalty is defined as explicitly penalizing the trace of the Fisher information matrix (FIM). Jastrzebski et al. (2021), Stickland and Murray (2021) observed penalizing FIM during training correlates to better generalization performance. It can be written mathematically as $\frac{1}{n}\sum_i^n \nabla L(x_i, y_i)$ where $L(x_i, y_i)$ is the loss at the data point (x_i, y_i).

MVR (Wang et al., 2021) function on the concept of consistency regularization where the divergence between the model predictions on deterministic and probabilistic segmentation inputs is minimized. The objective function is formulated as

$$\sum_{i=1}^{N}\Big(-\frac{1}{2}\log p(y_i|\hat{x}_i) - \frac{1}{2}\log p(y_i|x'_i) \quad (1)$$

$$+ \lambda D(p(y_i|\hat{x}_i) \parallel p(y_i|x'_i))\Big) \quad (2)$$

where the first term is the model loss on *deterministic segmentation* of the i^{th} data sample (most probable segmentation), the second term is the model loss on *probabilistic segmentation* of the i^{th} data sample (random segmentation) and the third term is the KL divergence between these two output predictions. This technique influences the model to be consistent on the predictions of different input types which successively motivates the model to be more adversarially robust.

3.2 Generalization Measures

Our study aims to investigate which type and characteristics of methods would best correlate with better performance in generalization, in this case, zero-shot cross-lingual transfer. We are especially interested in confirming the applicability of the flatness hypothesis for cross-lingual generalization. In order to assess whether a flat optimum loss-scape region corresponds to generalization, we essentially break down the experiment to measure two things, flatness, and generalization, such that their correlation can be measured.

Jiang et al. (2020b) conducted an extensive study on image classification tasks using generalization measures such as flatness-based measures (sharpness metrics), margin and norm-based metrics (based on parameter norms and distance from initial weights) to find correlations between measures and model performance which supported the usability of measures. These measures can be useful to explore the capabilities of language models to transfer knowledge from high-resource languages to low-resource ones.

Margin

Higher certainty in predicting the correct label leads to a model that is robust to perturbations and unseen examples. Margin is the distinction between model prediction for ground truth label and the next highest prediction probability. We use an average based margin formula defined by Wei et al. (2018) to calculate margin values on the entire test set. Jiang et al. (2020b) observed that the margin was directly proportional to better generalization in the image classification tasks. Margin is

$$\frac{1}{n}\sum_{i}^{n}\Big(f_{y_i}(x_i) - \max_{j\neq y_i} f_j(x_i)\Big)$$

where x_i is the i^{th} input to model, y_i is the ground truth label, $f(.)$ is the model function. A larger value of the margin of a model on a given dataset would mean higher confidence in the model to predict the correct label - including unseen examples (from languages not included in fine-tuning).

Sharpness of optimum

In simpler terms, we can define sharpness as the change in the model loss value at two neighboring points in the model weights plane. It can also be loosely interpreted as the inverse of the maximum radius the loss function can sustain a low loss value at the optimum. Sharpness-based measures resulted in the highest correlation with generalization in (Jiang et al., 2020b).

Jiang et al. (2020b) formulates the sharpness to be

$$\phi = \frac{\|W - W_0\|_2^2 \log(2\omega)}{4\alpha^2} + \log\frac{m}{\sigma} + 10$$

such that $\max_{|u_i|\leq\alpha} L(f_{W+u}) < 0.1$, where α is the maximum radius in the model's loss landscape possible, W and W_0 are the models finetuned weights and model initial weights respectively, ω is the number of parameters, m is the total number of observations, σ is the standard deviation of Gaussian noise added. In this work, as we are comparing models with the same architecture (considering mBERT only), on the same dataset, we can remove the constants, and simplify the equation further for comparative analysis.

$$\phi = \frac{\|W - W_0\|_2^2}{4\alpha^2}$$

Intuitively, if the radius of the low-loss region in the loss-landscape (α) is small, that means the model has a higher loss value near the optimum, which would mean the landscape of the optimum is not flat. We can relate this to resulting in an unstable prediction when having perturbations in either the data or model weights. Jiang et al.'s formula didn't result in stable results for our experimental set-up which might be because the ascent steps taken to optimize the α value resulted in either having a large or a very small final α. The values of α occurred at extreme points because the algorithm was using a binary search method and whenever optimal α was not found, the search algorithm stopped with the final α value at either of the extreme points. The correlation results of the above sharpness method are shown in Table 3.

We present an alternative definition (inclined with sharpness measure mentioned in the works of Keskar et al., and Foret et al.), $\phi_{\text{difference}}$ that removes the need to optimize α by calculating the difference between loss values at two points in the optimum region, formulated as

$$\phi_{\text{difference}} = L(f_{W'}) - L(f_W)$$

where W' is $W + \epsilon$ (ϵ being Gaussian noise) and W is the optimum weight parameters. The details of our definition are in Algorithm 1 and performs calculation at about roughly 5-10 times faster than Jiang et al.'s algorithm for a given batch size of 8.

Algorithm 1 Difference-based sharpness algorithm

1: w_0 = original_weight
2: $w = w_0 + \epsilon$ ▷ Small noise added to avoid zero gradient
3: $\Delta w = \nabla L(w)$
4: $w' = w + n\Delta w$
5: $p = \lambda \times \|w'\|_F$ ▷ λ is small like 0.05
6: **if** $\|w' - w_0\| > p$ **then**
7: $\quad w' = w_0 + \frac{(w' - w_0)}{\|w' - w_0\|} \cdot p$
8: **end if**
9: $\phi_{\text{difference}} = L(w') - L(w_0)$

4 Experiments

For comparison, we implement each Sharpness-based optimization as a fine-tuning objective on the multilingual mBERT base variant (`bert-base-multilingual-cased` from huggingface) (Devlin et al., 2019) in addition to mT5 model (`google/mt5-small`) (Xue et al., 2021). We use a linear classification layer of size 768x3 where the output dimension is equal to the number of labels. We adopt a two-step training approach in our experiments. First, we fine-tune the model on the English language part of the XNLI dataset to optimize the model to learn the task specifically in English. Subsequently, we perform a zero-shot transfer of the fine-tuned model on the rest of the 14 languages to facilitate an evaluation of the generalization of models.

4.1 Data, Model details, and Settings

For this work, we used the XNLI dataset (Conneau et al., 2018) that includes data samples from the MultiNLI dataset (Williams et al., 2018) and their translated versions in 14 different languages (Arabic "ar", Bulgarian "bg", German "de", Greek "el", Spanish "es", French "fr", Hindi "hi", Russian "ru", Swahili "sw", Thai "th", Turkish "tr", Urdu "ur", Vietnamese "vi", Chinese "zh"). We only train the models on the English ("en") subset of the dataset. We use the data of these 14 languages only for inference and evaluation of the models.

We fine-tune pretrained mBERT models for 15 epochs each with a batch size of 32, with a learning rate of 2×10^{-5}, and select best checkpoint on validation. The objective function we use for the baseline model with the classification layer is the AdamW optimizer (Loshchilov and Hutter, 2019) with cross-entropy loss, the mBERT+FIM model has an additional loss as Fisher Penalty, the mBERT+SAM model uses the SAM optimizer and mBERT+MVR uses the MVR algorithm for fine-tuning. We use the hyperparameters and code presented in XTREME[2] and MVR codebase[3]. We run the models with 8 random seeds and present the average performance of these models (Figure 6). In Algorithm 1, the amount of Gaussian noise we add to model weights during calculating sharpness is controlled using a scale that we empirically find (among [0.001, 0.005, 0.01, 0.02]) for each model, with n equal to 0.05.

We fine-tuned the MT5 model (`google/mt5-small` using Huggingface's library over 15 epochs. The XNLI dataset was

[2]https://github.com/google-research/xtreme

[3]https://github.com/cindyxinyiwang/multiview-subword-regularization

processed using a function to tokenize inputs, and the optimizer utilized was Adafactor with a learning rate scheduler. Adafactor optimizer's ability to adapt learning rates is helpful with larger models like T5 in multi-lingual settings. We trained the model with a batch size of 8, accumulating gradients over 4 steps.

Additional experiments were run on PAWS-X dataset (Yang et al., 2019) which has 7 languages: German "de", English "en", Spanish "es", French "fr", Japanese "ja", Korean "ko", Chinese "zh". We use similar experimentation of fine-tuning on english and doing a zero-shot transfer on 6 other languages as defined above for this dataset. We used Huggingface's models: mBERT (`bert-base-multilingual-cased`), RoBERTa (`roberta-base`), and XLM (`xlm-mlm-en-2048`) using Adam optimizers.

Results

To evaluate how each of the selected measures correlates with cross-lingual generalization, we first compare these measures with held-out test accuracy. In Table 1, we present the correlation coefficients (using `numpy.corrcoef`) of margin vs. accuracy and sharpness vs. accuracy. We notice that having a higher margin is exceptionally correlated to achieving great performance on unseen language data. Hence, we assume the margin to indicate the generalization performance of a given model. Similarly, sharpness captures a noteworthy negative correlation with test performance.

Model	Correlation with Accuracy	
	Margin	Sharpness
Baseline	0.801	-0.845
mBERT+MVR	0.818	-0.793
mBERT+SAM	0.874	-0.584
mBERT+FIM	0.954	-0.671
mT5 + Adafactor	0.912	-0.410

Table 1: Correlation coefficients between Margin & Test Accuracy, and Sharpness & Test Accuracy on the XNLI dataset.

We notice similar results by extending our similar experimentation to Paraphrase Identification, PAWS-X dataset (Yang et al., 2019) with 3 different models: mBERT (`bert-base-multilingual-cased`), RoBERTa (`roberta-base`) (Liu et al., 2019), and XLM (`xlm-mlm-en-2048`) (CONNEAU and Lample, 2019) and analyze the validity of the flatness hypothesis, i.e. a flat optimum neighborhood would lead to a generalized model. In Figure 1, we confirm the strong relationship between Margin (indicating generalization) and Sharpness (indicating flatness) even when compared across all models and metrics, suggesting flat neighborhoods of model optimum can help in achieving higher margin values which correlate to better generalization. More findings about visualizations are in Appendix A.1.

Model	Correlation with Accuracy	
	Margin	Sharpness
mBERT	0.998	-0.289
RoBERTa	0.997	-0.708
XLM-R	0.995	-0.622

Table 2: Correlation coefficients between Margin and Sharpness with Test Accuracy on the PAWS-X dataset.

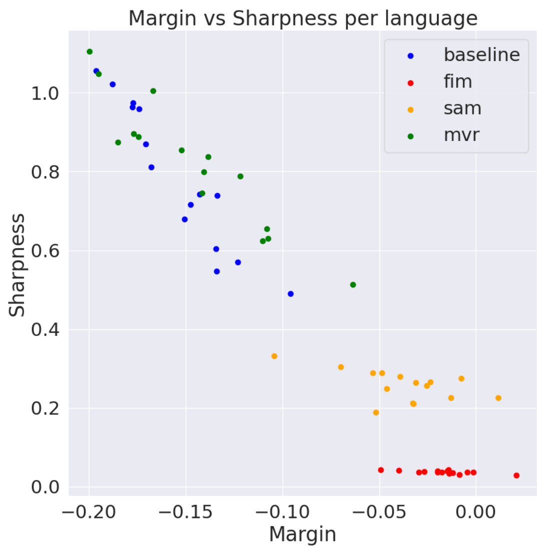


Figure 1: Scatter plot of Margin values and Sharpness ($\phi_{\text{difference}}$) values for each mBERT model (on XNLI dataset) with different objectives language-wise to show the relationship between sharpness and generalization.

We can interpret sharpness as the inverse of flatness, providing us the verdict that flatness of the minimum in which the fine-tuned model is, would help the model perform better on unseen language data. When we evaluate similar models trained with different objectives across languages, we ob-

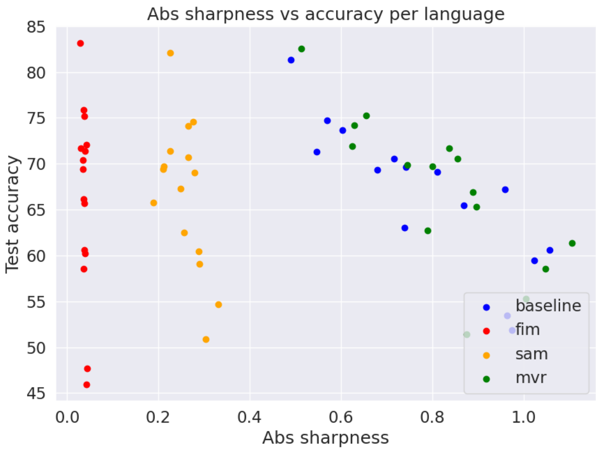


Figure 2: Scatter plot of difference-based sharpness measure with test performance for all models combined.

serve that the relationships between measures are likely dependent on the optimization objective functions used during fine-tuning. In coherence with both Figure 1 and 2, overall, we see that min-max based optimization methods including FIM and SAM, have the lowest sharpness values, compared to the baseline and the regularization method MVR.

In Figure 3, we create scatter plots for mBERT models where in each scatter plot, we plot the model's margin based on the validation set for each language, and we plot the accuracy of that model on the test set on the XNLI dataset. We observe that the margin measure exhibits a consistent correlation with test performance across all the models analyzed.

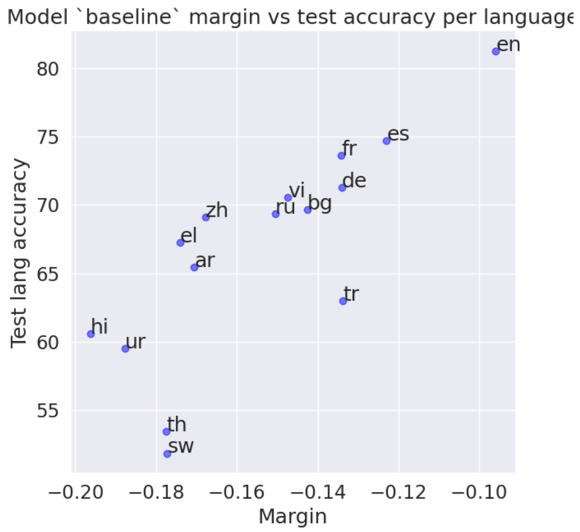
As can be seen in the scatter plots for sharpness (proposed difference-based sharpness) and accuracy in Figure 4, findings further indicate a negative correlation between sharpness and test performance, suggesting that lower sharpness values are associated with better generalization, represented as model performance on unseen data.

5 Conclusion

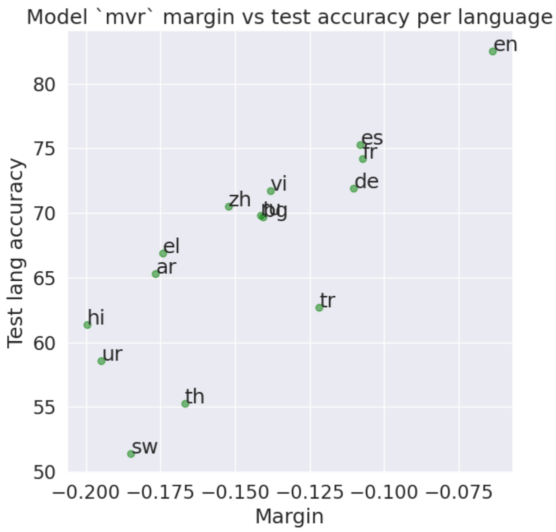
Enabling cross-lingual knowledge transfer is an important step towards extending the applicability of NLP models to more languages. Despite recent efforts to develop better optimization methods for improving the generalization of language models in new languages or domains; these techniques try different types of methods to achieve higher performance such as sharpness-based minimizations, reducing gradient of loss functions, or consistency regularization. Evaluating these techniques thoroughly without a standardized methodology remains a difficult task. This work aims to uncover insights into how to measure cross-lingual generalization by exploring suitable measures that work well under different settings. Our experiments studying model loss landscape and parameter properties find strong relationships between the margin, sharpness in the loss minima neighborhood, and zero-shot cross-lingual downstream task performance, both on validation and test sets, supporting strong applicability to evaluate models before deploying them in new languages.

Limitations

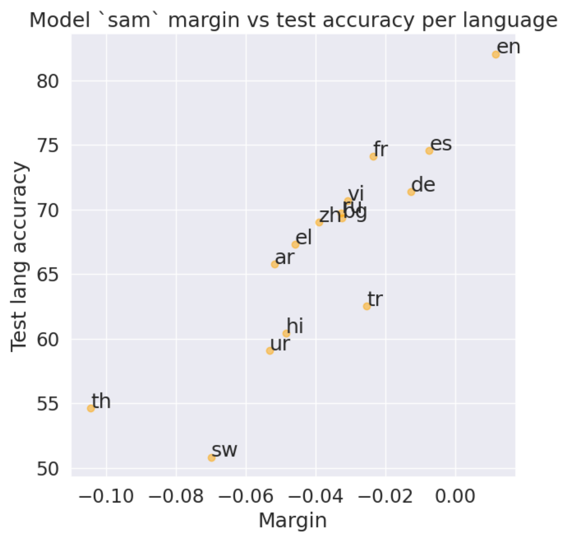
The algorithm presented in our paper, the difference-based sharpness measure, is a great novelty for more robust sharpness computation, however, we would like to acknowledge that a few variables in the algorithm still require tuning heuristically, including the noise scale and the multiplication coefficient required to compute the projected radius. Secondly, the mean-based margin distance is only applicable to classification tasks. Due to the limited scope of this project, we leave the development of generalization measures more suitable for generative tasks to future work.



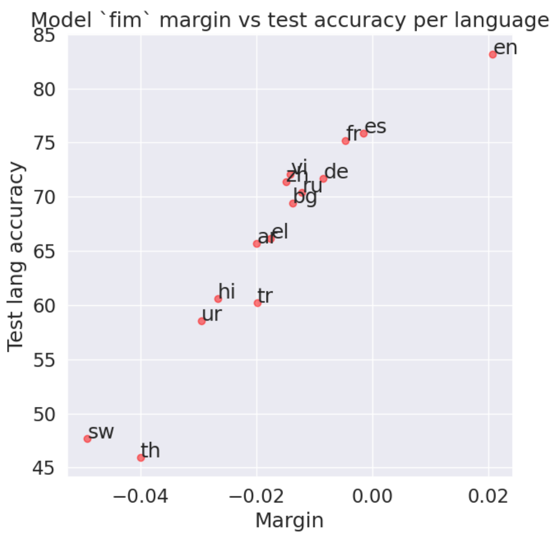
(a) Correlation visualization for Baseline(mBERT + AdamW)



(b) Correlation visualization for mBERT + MVR model

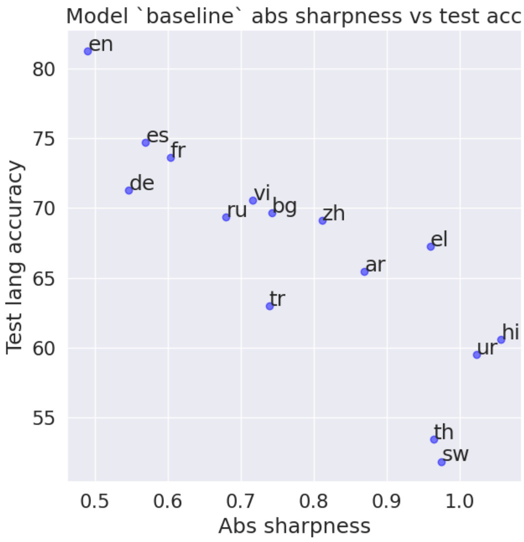


(c) Correlation visualization for mBERT + SAM model

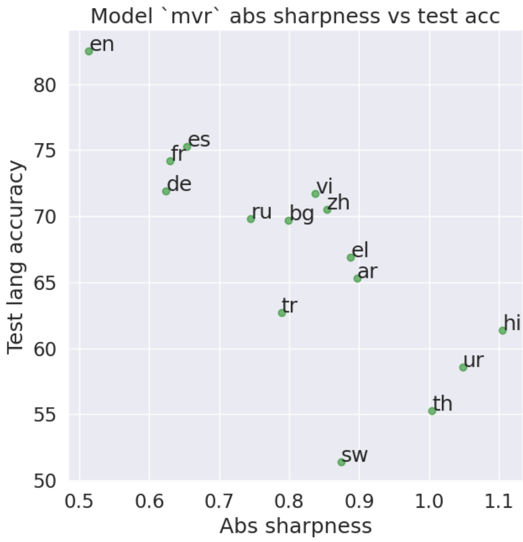


(d) Correlation visualization for mBERT + FIM model

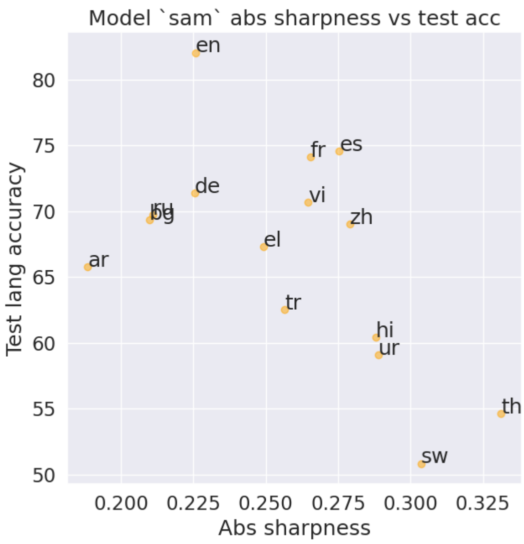
Figure 3: Scatter plots of margin of individual models and their corresponding performance on test set language-wise on XNLI dataset.



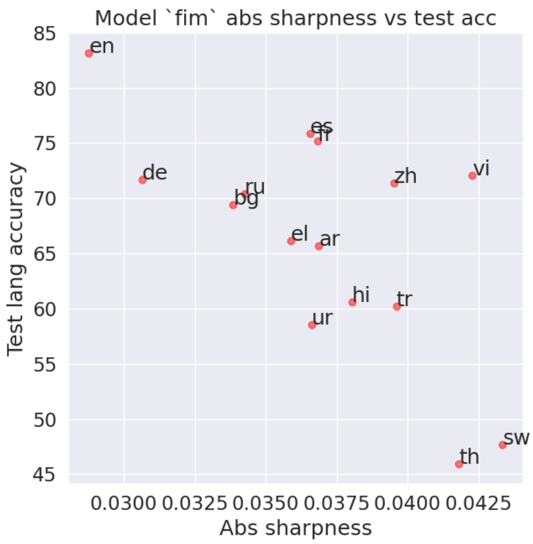
(a) Correlation visualization for Baseline (mBERT + AdamW)



(b) Correlation visualization for mBERT + MVR model



(c) Correlation visualization for mBERT + SAM model



(d) Correlation visualization for mBERT + FIM model

Figure 4: Scatter plots of the proposed difference-based sharpness ($\phi_{\text{difference}}$) of individual models and their corresponding performance on test set language-wise on XNLI Dataset.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Léon Bottou. 2012. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. 2019. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, (12):124018.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. 2017. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR.

Erik Englesson and Hossein Azizpour. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297.

Ronald Aylmer Fisher. 1925. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hang Hua, Xingjian Li, Dejing Dou, Chengzhong Xu, and Jiebo Luo. 2021. Noise stability regularization for improving bert fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3229–3241.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. 2021. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4772–4784. PMLR.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020a. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020b. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. 2022. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Lijun Liang, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. The MIT Press.

Ido Nachum, Jonathan Shafer, Thomas Weinberger, and Michael Gastpar. 2024. Fantastic generalization measures are nowhere to be found. In *The Twelfth International Conference on Learning Representations*.

Vaishnavh Nagarajan and J. Zico Kolter. 2019. Generalization in deep networks: The role of distance from initialization. *arXiv:1901.01672 [cs.LG]*.

Jungsoo Park, Gyuwan Kim, and Jaewoo Kang. 2022. Consistency training with virtual adversarial discrete perturbation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5646–5656.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. 2021. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432.

Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Asa Cooper Stickland and Iain Murray. 2021. Regularising fisher information improves cross-lingual generalisation. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 238–241.

V. N. Vapnik and A. Ya. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. 2018. On the margin theory of feedforward neural networks. *arXiv:1810.05369 [stat.ML]*.

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)*, pages 1112–1122. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417.

A Appendix

A.1 Visualization results

Previous work (Nagarajan and Kolter, 2019; Jiang et al., 2020b) suggests that a lower Frobenius distance from initialization would lead to better generalization. As Figure 5 shows, we fail to observe a strong direct relationship between generalization and Frobenius distance from initialization. However, the model trained with Fisher Penalty as an additional objective function that has a high distance from initialization overall performed poorly than others. We also see that models trained with Fisher Penalty, SAM, and MVR optimizers tend to be more stable than the baseline model, with Fisher Penalty resulting in the most stable model when trained multiple times (with different seeds, see Figure 6), and SAM achieving generally the best average zero-shot task accuracy.

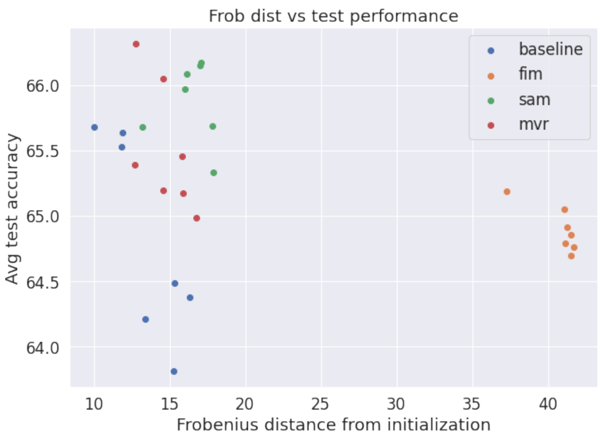


Figure 5: Scatter plot of Frobenius distance from initialization and Test accuracy for each model type (trained multiple times independenty).

A.2 Additional experiments

To compare all models together, using the difference-based sharpness measure, on language-wise performance, we observe it is dependent on the learning algorithms used during training in Figure 2.

We used the Jiang et al.'s α-based sharpness algorithm for the experiment and optimized the threshold loss values for our experimental setting. The results of the correlation coefficient (using `numpy.corrcoef`) for α-based sharpness and test accuracy are shown in Table 3 and Figure 7. We notice that α-based sharpness values occur at extreme points (for example, for mBERT+FIM model, sharpness values are low whereas for the

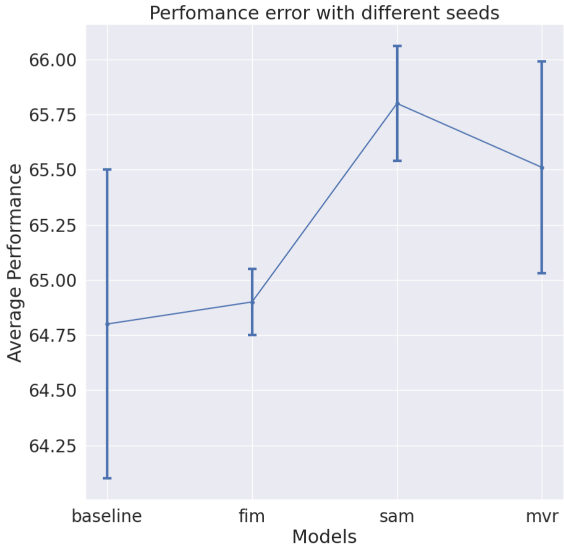
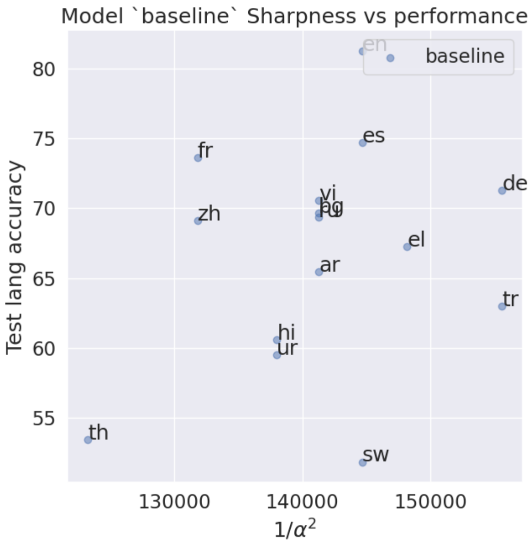


Figure 6: Average test performance (and deviations) of models when trained multiple times with different seeds.

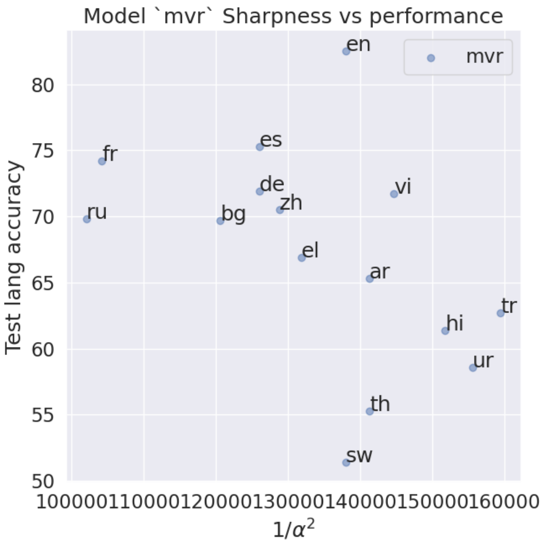
Model	Correlation coefficient of α-sharpness with accuracy
Baseline	0.249
mBERT+MVR	-0.471
mBERT+SAM	-0.166
mBERT+FIM	-0.440

Table 3: Correlation coefficients between α-sharpness (Jiang et al., 2020b) & Test Accuracy on the XNLI dataset.

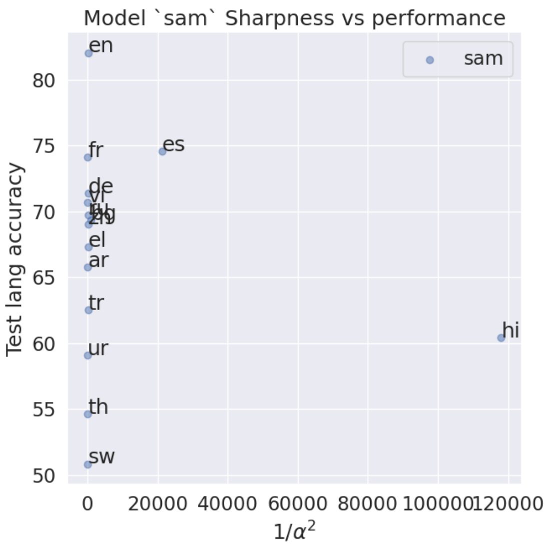
Baseline or mBERT+MVR model, sharpness values are much larger). Apart from being a computationally expensive algorithm, we failed to see a strong relationship of α-based sharpness with performance in Baseline and mBERT+SAM models.



(a) Correlation visualization for Baseline (mBERT + AdamW)



(b) Correlation visualization for mBERT + MVR model



(c) Correlation visualization for mBERT + SAM model

(d) Correlation visualization for mBERT + FIM model

Figure 7: Scatter plots of Jiang et al. (2020b) based α-sharpness measure (we are only considering $\frac{1}{\alpha^2}$ here) of individual models and their corresponding performance on test set language-wise.