Towards Detecting, Augmenting, and Sampling, Efficient Datapoints for Robust Out-of-Distribution Generalization

Atharva Kulkarni, Raghav Kapoor, Anubha Kabra
Carnegie Mellon University, Language Technologies Institute
{atharvak, raghavka, anubhak}@cs.cmu.edu

Abstract

As most machine learning models are trained and evaluated on datasets that are independent and identically distributede (i.i.d), their performance falters when faced with samples that are adversarial or statistically out-of-distributed (OOD). Thus, in this project, we work towards robust out-of-distribution generalization without having access to OOD samples. Specifically, we propose a threefold approach for training deep neural networks. To begin with, we identify samples that are easy to learn, mislabeled, and hard to learn using the model's training dynamics. Then we proffer a mixup-based data augmentation technique that generates effective synthetic samples intelligently combining easy and difficult samples that contribute towards improving model robustness and generalization capabilities. Lastly, we plan on adopting a non-trivial sampling scheme that makes the learning procedure more efficient. In this report, we begin by experimenting with detecting and calibrating large language models for out-of-distribution detection by reproducing experiments of Arora et al. (2021).

1 Introduction

Modern-day NLP is characterized by large datasets, and big benchmarks (Bowman et al., 2015; Rajpurkar et al., 2016; Wang et al., 2018). While many of these corpora are manually curated through human annotation, others use weakly-supervised labeling schemes to circumvent the tedious and financially expensive annotation process (Malik and Bhardwaj, 2011). Though increasing the amount of labeled data improves the likelihood of learning diverse data distributions, the data quality plays a significant role in robustness and out-of-distribution generalization. Whether it be automatic labeling or annotation bias, almost all current datasets are interspersed with duplicate, untidy, redundant, mislabelled, corrupted, or ambiguous examples hampering the quality of the dataset. Such label noise can hinder the generalization capabilities of the over-parameterized neural networks, owning to their strong proclivity to memorization (Arpit et al., 2017; Zhang et al., 2021). Therefore, assessing dataset quality and finding examples that contribute to achieving good performance on in-distribution as well as out-of-distribution samples is an open research problem.

While ambiguity is intrinsic to natural language, it is vital to identify OOD data samples (Arora et al., 2021) and assess which data points contribute to better learning and generalization (Swayamdipta et al., 2020; Karamcheti et al., 2021). Moreover, training on just good samples is not enough; it is essential to include examples from diverse distributions to help generalize better on in-domain as well as out-of-distribution data points (Linzen, 2020). Lastly, developing an efficient sampling scheme is critical to how a deep neural network learns different features without incorporating spurious correlations and label memorization. Evidence has shown that non-typical sampling schemes that focus on example importance and difficulty yield better overall and generalization performance (Sachan and Xing, 2016; Liu et al., 2018; Xu et al., 2020a; Peng et al., 2022b,a).

To this end, we propose a novel methodology encapsulating the three aforementioned points of view for robust natural language understanding. To begin with, we look at different techniques to identify OOD samples, and how to assess models on their OOD performance (Arora et al., 2021). Next, we propose a training dynamics-based approach that identifies easy-to-learn, ambiguous, and hard-to-learn examples (Swayamdipta et al., 2020). This spawns examples that contribute positively towards fast learning, examples that help generalize, and mislabelled examples, respectively. To widen the distribution of learn-worthy examples, we suggest a mix-up strategy that uses a combination of easy-to-learn and ambiguous groups to produce high-

quality examples. Finally, taking insights from the works of Sachan and Xing (2016), Liu et al. (2018), (Guo et al., 2020), (Xu et al., 2020a), and Mindermann et al. (2022), we plan on developing a non-typical data sampling scheme that focuses on iteratively learning samples in the increasing order of difficulty. We hypothesize that our three-fold training scheme would lead to better overall and generalization performance without the need for any fancy optimization or auxiliary data.

In this report, we provide a thorough literature review of the research that has been done in relation to our problem statement. We also experiment with our first approach to identifying OOD examples and assessing the model's capabilities on their OOD performance. With text classification as a case study, we reproduce the results by Arora et al. (2021) and conduct a thorough detailed error analysis. Based on its insights, we posit improvements and our proposed methodology.

2 Literature Survey

Detecting OOD examples has become an important research topic. Hendrycks et al. (2020) show that pre-trained transformers are much more efficient at detecting anomalous or OOD examples compared to their other neural network counterparts. Kong et al. (2020) proposed *on-manifold regularization* and *off-manifold regularization* to train transformers for improved in-distribution and out-of-distribution calibration, respectively. Zhou et al. (2021) posit an unsupervised OOD detection technique with a contrastive-learning-based training scheme where they fine-tune transformers on ID data. Arora et al. (2021) attest that density estimation methods are more suitable for background shift settings, while the calibration methods perform well for datasets with semantic shifts.

Training dynamics-based approaches are popular in machine learning literature to assess dataset difficulty and quality. Pleiss et al. (2020) proposed Area Under the Margin (AUM) statistic to distinguish correct vs. mislabelled samples. It measures the average difference between the logit values for an example's assigned label and its highest non-assigned label. Correctly labeled examples exhibit higher AUM values, whereas the hard and mislabelled examples are identified using an empirical threshold. Swayamdipta et al. (2020) extends this idea to identify easy-to-learn, ambiguous, and hard-to-learn examples. For each example, they calculate confidence and variability based on the mean and standard deviation of the gold label probabilities. The easy-to-learn, ambiguous, and hard-to-learn examples are categorized as examples with high confidence + low variability, high variability, and low confidence + low variability, respectively. Toneva et al. (2019) utilized training dynamics to identify frequently misclassified examples during later training epochs, despite being classified correctly during initial epochs. Their experiments reveal interesting findings: (i) specific examples are forgotten with high frequency, whereas some are never forgotten; (ii) The un-forgettable examples generalize across neural architectures; and (iii) A significant fraction of examples do not contribute much to the model's learnability and can be omitted during training while still maintaining state-of-the-art generalization performance. Studying training dynamics through the lens of identifying important examples, Paul et al. (2021) present two novel metrics, namely Gradient Normed (GraNd) and the Error L2-Norm (EL2N) that successfully identify important and difficult examples early in training. They further attest that pruning the training dataset with small GraNd does not sacrifice test performance. Adila and Kang (2022) use training dynamics to identify out-of-distribution samples.

More recently, Chiang and Lee (2022) investigated input representations, hidden layer representations, and output probabilities to distinguish in-distribution (ID), out-of-distribution (OOD), and adversarial examples. Their findings reveal that OOD examples exhibit differences from the initial layers themselves, while adversarial examples yield output probabilities with low confidence. Xu et al. (2021) provide an effective transformer approach to identify OOD examples using unsupervised in-domain data. On another note, Adila and Kang (2022) focuses on identifying dataset difficulty and difficult examples based on V-usable information (Xu et al., 2020b). They present pointwise V-information (PVI) for measuring the difficulty of individual instances w.r.t. a given distribution. Finally, Mindermann et al. (2022) prioritize finding examples that are learnable, worth learning and yet to learn using a probabilistic modeling selection function.

Mixup (Zhang et al., 2018) is a popular technique for data augmentation where additional samples are generated using a combination of different training examples and their labels. Sun et al.

(2020) proposed a Mixup-Transformer with a stack of a mixup layer over the final hidden layer of the pre-trained transformer-based model to do hidden space interpolation. Yin et al. (2021) come up with Batch-Mixup, which interpolates hidden states of the entire mini-batch, generating new points scattered throughout the space of the corresponding mini-batch. (Guo et al., 2019) presented a study of mixup-based interpolation of the word and sentence embeddings for text classification. Guo (2020) proffered a novel non-linear mixup technique where the mixing policy for the labels is adaptively learned based on the mixed input.

Machine learning literature reflects that techniques such as curriculum learning (Wang et al., 2022) and importance sampling (Li et al., 2020) that sample examples in a non-randomized fashion result in improved performance. Initial work by Sachan and Xing (2016) showed that incorporating an easy yet diverse set of samples via self-paced learning can yield improvements for non-convex VQA models. Xu et al. (2020a) come up with a similar approach to natural language understanding. Liu et al. (2018) proposed a curriculum learning-based framework for natural answer generation (CL-NAG), which initially samples simple and low-quality QApairs and then incorporates more complex and higher-quality ones, resulting in a model that gradually learns to produce better answers with richer contents and more complete syntaxes. Guo et al. (2020) on the other hand, present a curriculum learning-based approach that progressively switches the training from training a machine translation model from autoregressive generation to nonautoregressive generation.

Drawing insights from these three approaches, we propose a unified framework that identifies a different variety of data points, does intelligent mixup, and learns them in a curriculum-learning fashion.

3 Baseline Implementation

In this report, we reproduce the experiments done by Arora et al. (2021). To detect OOD examples, Arora et al. (2021) use two methods. The calibration method leverages a model's confidence score as the final scoring function and indicates the probability of the model's capacity to predict correctly. We generally use conditional probabilities from the model which is useful for fine-tuning pmodel from pre-trained other models. The second technique is Density Estimation, wherein the density estimator is used to predict the likelihood of a particular instance which is used as the score function. Using these two scoring functions, the experiments are performed on multiple pairs of ID/OOD dataset combinations and categorized as either semantic shift or background shift. Background shift relates to those instances when the language pattern or word usage is similar in nature and what changes in the context where that language is used. For example, if we consider two datasets, one with a movie review and another with online shopping product reviews, then both have a similar connotation of the words in the review, like "This movie is great" *versus* "This product is great". However, only the context or the subject changes in both of these scenarios, while the usage pattern remains the same. In our set of experiments, we consider only datasets and experiments related to background shift as they would form the baselines of our proposed work for future tasks.

3.1 Re-Implementation and Training Details

As a part of the re-implementation task, we recreate a similar setup as the authors (Arora et al., 2021). Out of all the experiments that are performed in the paper, we meticulously select a sample that aligns with our future direction of work. For these experiments, we train the model mainly on RoBERTa Base (with 110M parameters) is used, which helps to control the size of the model. Also, we choose three datasets - Yelp, SST2, and IMDB, all of which relate to the task of sentiment analysis, and try out all possible pairwise combinations of these three datasets. One of these datasets is considered ID while the other is OOD. The RoBERTa base model is run for 3 epochs for each of the dataset pairs, with a learning rate of 1e-5 and a batch size of 16. The OOD detectors are assessed on AUROC metrics. As the starter code, we use the code-base[1] provided by the authors, on top of which we add our changes to automate the scripts and we add additional classes to implement models for the Yelp dataset. Also, while fine-tuning we add some more features to record the metrics and final results. The updated repository can be found at[2]. For the purpose of training the models, we use AWS p2.xlarge boxes to speed up our computations. Also, we have provided a script, which can be used to run the

[1] https://github.com/uditarora/ood-text-emnlp
[2] https://github.com/anubhakabra/OOD-Generalization

Dataset	OOD	O-Accuracy	R-Accuracy	O-AUCROC	R-AUCROC	O-ID	R-ID
SST2	IMDB	92.0	91.79	66.2	68.68	93.8	94.15
	YELP	94.4	95.01	57.5	65.19		
IMDB	SST2	89.2	89.33	82.6	80.32	95.5	95.39
	YELP	95.4	94.86	67.1	73.40		
YELP	SST2	88.9	87.17	85.9	87.54	98.2	97.89
	IMDB	93.2	92.66	61.8	70.95		

Table 1: This table represents the original (O) and reproduced (R) results. ID: In Domain Accuracy.

Dataset	Train	Validation	Test	Usecase
SST2	67349	872	1821	Movie review
IMDB	25000	25000	50000	Movie review
Yelp_Polarity	560000	-	38000	Business Review

Table 2: Dataset Analysis

experiments. We could successfully reproduce the original results, which are similar to the author's original scores. This is discussed in Section 4.

4 Results and Discussions

Table 1 shows the results for our semantic shift pairs in comparison with the original paper (Arora et al., 2021). We have achieved replication of the baseline results for all the scenarios applicable. The results obtained are $\pm 0.5 - 1\%$ different from original results for most cases.

The *SST2* and *IMDB* datasets consist of movie reviews with different lengths. Meanwhile, the *Yelp_Polarity* dataset contains reviews for different businesses, representing a domain shift from *SST2* and *IMDB*. Each of these datasets is used as ID/OOD, using the validation split of *SST2* and test split of *IMDB* and *Yelp_Polarity* for evaluation.

We see a jump in reproduced AUC-ROC score for the *Yelp_Polarity* dataset as compared to the original scores. As mentioned in Section 4, the codebase did not include the Yelp_Polarity dataset for experimentation, although the base paper provides experimentation on this data. Hence, we made changes in the code base for experimentation on this dataset. While we kept all the hyperparameters as mentioned in the base paper, we see a boost in the AUC-ROC score for this data with respect to both out-of-data distributions from *SST2* and *IMDB* data.

Figure 1, 2, 3 shows the variation of AUC-ROC curve for IMDB, SST2, Yelp_Polarity data respectively with respect to different OOD data. We observe that both IMDB and Yelp_Polarity datasets

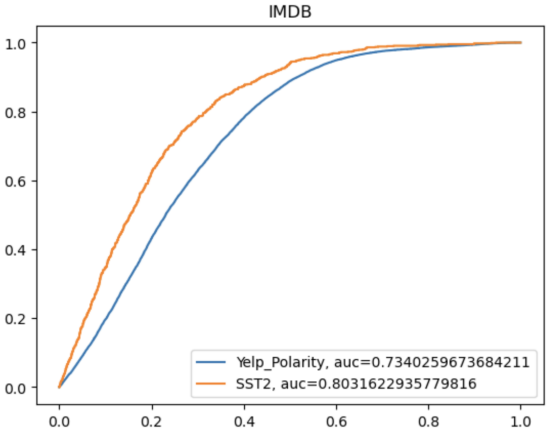

Figure 1: AUC-ROC curve for IMDB dataset replication. The OOD data used are SST2 and Yelp_Polarity respectively.

upon training are able to give a high AUC for the SST2 dataset. Using SST2 for training, however, does not translate to as good results on the IMDB and Yelp_Polarity data, as the curve flattens eventually. From table 2, we see that SST2 has the least training data relatively, hence the other datasets when used for training give better performance.

5 Error Analysis

The above implementation performs well for the sentiment analysis task for all the pairs of datasets that we considered. However, we still believe that there is significant scope for improvement that can be done to boost the scores. To understand which approaches to leverage, we must first focus on the current shortcomings of the methodologies as proposed by the current paper (Arora et al., 2021).

For sentences that have repeating phrases, the premise and hypothesis tend to overlap with each other and this results in confusing the model, inhibiting its capability to learn, which results in the model underperforming for such data points. Another issue associated with premise-hypothesis

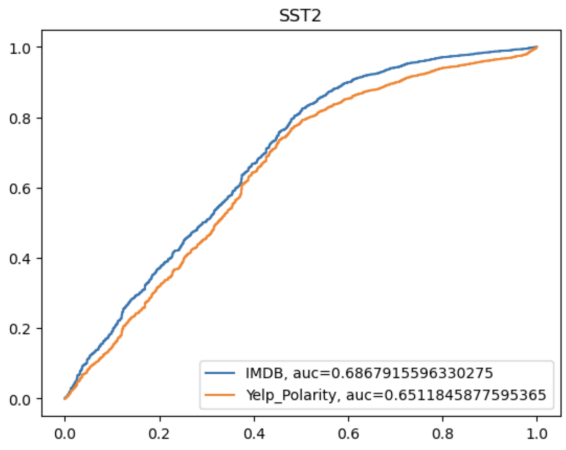

Figure 2: AUC-ROC curve for SST2 dataset replication. The OOD data used are IMDB and Yelp_Polarity respectively.

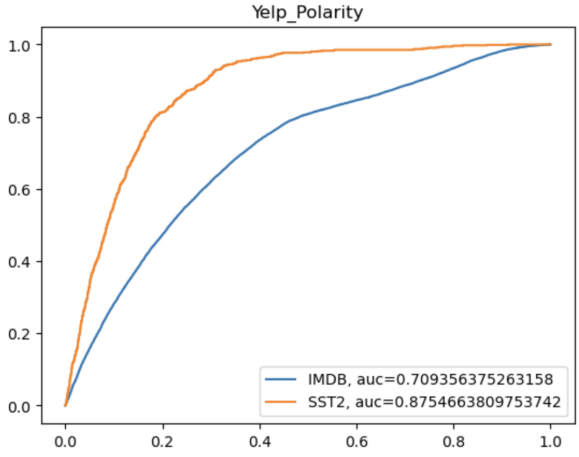

Figure 3: AUC-ROC curve for Yelp_Polarity dataset replication. The OOD data used are IMDB and SST2 respectively.

overlap is that model tends to learn a bunch of spurious features when run on ID data points, which forces it to perform much worse for OOD data, where it predicts the wrong answer with great confidence. This implies that the model could not generalize from the ID distributions even with a greater number of data points. As we discuss in the next section, this problem can be remediated with the intelligent mixup technique, where we augment a good mix of both easy-to-learn (ID) and hard-to-learn examples (OOD) to make the model more robust to such semantic feature overlaps.

Also, the model fails to detect any small shifts in the background across the OOD dataset in many cases, especially when the overlap in distributions is significant. Due to this, we see a worsened performance for OOD. The model as we see is particularly invariant to small background shifts rather than semantic shifts. To make the model capable of detecting such subtle shifts in the background contexts across different kinds of datasets, we could also train our models via the curriculum learning technique, wherein we make the model learn easy examples first before training them further on the harder examples. This way, the model is powerful enough to detect these nuances in the background, which it might fail to decipher if random sampling is done.

In general, we see that the AUCROC measure is pretty much stable across the experiments. However, there seems potential to boost these scores using the techniques discussed in the next section.

6 Proposed Framework

In this section, we briefly elaborate on our proposed three-fold approach. Combining ideas from three different standpoints, we propose a robust pipeline for different NLP tasks that is dataset as well as task agnostic. In our work, we plan to experiment with text classification.

Data Quality Estimation: Reproducing the experiments by Arora et al. (2021) we find that language models struggle to preform well on OOD data. Thus, in our future work we propose to mine samples that are easy to learn and hard to learn, and differentiate them from the ambiguous and potentially mislabeled ones, building on the work of Swayamdipta et al. (2020). We hypothesize that that learning with the pruned dataset containing easy-to-learn and hard-to-learn examples leads to better performance than training on the entire dataset. We use this as our backbone model for further investigations.

Intelligent Mixup: Data augmentation based on Mixup strategies has been shown to boost both the generalization and robustness of deep neural networks (Zhang et al., 2018). While mixup takes a convex combination of pairs of examples and their labels in a randomized fashion, we propose a more intelligent technique. We plan to generate synthetic samples by interpolating specific combinations of examples, namely, (easy-to-learn, easy-to-learn), (easy-to-learn, ambiguous), and (ambiguous, ambiguous). We ignore the hard-to-learn examples as they often constitute mislabeled instances. More-

over, to encourage the generation of more diverse and challenging examples, we plan on combining semantically dissimilar examples as well as examples with high difficulty as measured by their PVI scores (Adila and Kang, 2022), respectively. Another approach that we are thinking of is doing interpolation in the hyperbolic space based on some recent works (Chen et al., 2022).

Curriculum Learning: Evidence in machine learning literature suggests that deciding the sampling scheme of data points apriori results in much better learning of neural networks than random sampling (Wang et al., 2022). Thus, we plan on an efficient sampling scheme that first learns easy examples, followed by the mixup-based synthetically generated easy instances. Progressively the difficulty of sampled examples increases with the next samples being the synthetic easy+ambiguous ones, original ambiguous, and finally the synthetic ambiguous examples.

7 Conclusion

Training models with good in-distribution and acceptable out-of-distribution performance is a difficult task. The task becomes even more challenging when we do not have access to OOD data samples. Thus, in this work, we propose to improve OOD performance via efficient data augmentation and training strategies. To begin with, we replicate experiments of Arora et al. (2021) to identify the performance of language models on OOD datasets. Focusing on text classification, we train models on one dataset and evaluate them on other datasets with either semantic or distribution shifts. Their calibration and density estimation methods suggest that the PLMs underperform under all conditions. Thus, to address this bottleneck, in our subsequent work, we plan a threefold strategy of finding learnworthy samples, doing intelligent mixups, and sampling them in a non-trivial fashion, the details of which are elaborated in this report. We are optimistic that our technique will yield performance gains compared to the standard training scheme.

References

Dyah Adila and Dongyeop Kang. 2022. Understanding out-of-distribution: A perspective of data dynamics. In *Proceedings on "I (Still) Can't Believe It's Not Better!" at NeurIPS 2021 Workshops*, volume 163 of *Proceedings of Machine Learning Research*, pages 1–8. PMLR.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Hui Chen, Wei Han, Diyi Yang, and Soujanya Poria. 2022. DoubleMix: Simple interpolation-based data augmentation for text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4622–4632, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2022. Understanding, detecting, and separating out-of-distribution samples and adversarial samples in text classification. *arXiv preprint arXiv:2204.04458*.

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4044–4051.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7839–7846.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. 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 7265–7281, Online. Association for Computational Linguistics.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum learning for natural answer generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4223–4229. International Joint Conferences on Artificial Intelligence Organization.

Hassan H. Malik and Vikas S. Bhardwaj. 2011. Automatic training data cleaning for text classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 442–449.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. 2022. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15630–15649. PMLR.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pages 20596–20607. Curran Associates, Inc.

Xinyu Peng, Fei-Yue Wang, and Li Li. 2022a. Towards better generalization of deep neural networks via non-typicality sampling scheme. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11.

Xinyu Peng, Jiawei Zhang, Fei-Yue Wang, and Li Li. 2022b. Drill the cork of information bottleneck by inputting the most important data. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6360–6372.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 453–463, Berlin, Germany. Association for Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020a. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

6095–6104, Online. Association for Computational Linguistics.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. 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 1052–1061, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020b. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Wenpeng Yin, Huan Wang, Jin Qu, and Caiming Xiong. 2021. BatchMixup: Improving training by interpolating hidden states of the entire mini-batch. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4908–4912, Online. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.