Understanding Out-of-distribution:A Perspective of Data Dynamics

Dyah Adila*
Department of Computer Sciences
University of Wisconsin-Madison
adila@wisc.edu

Dongyeop Kang
Department of Computer Science and Engineering
University of Minnesota
dongyeop@umn.edu

Abstract

Despite machine learning models' success in Natural Language Processing (NLP) tasks, predictions from these models frequently fail on out-of-distribution (OOD) samples. Prior works have focused on developing state-of-the-art methods for detecting OOD. The fundamental question of how OOD samples differ from in-distribution samples remains unanswered. This paper explores how data dynamics in training models can be used to understand the fundamental differences between OOD and in-distribution samples in extensive detail. We found that syntactic characteristics of the data samples that the model consistently predicts incorrectly in both OOD and in-distribution cases directly contradict each other. In addition, we observed preliminary evidence supporting the hypothesis that models are more likely to latch on trivial syntactic heuristics (e.g., overlap of words between two sentences) when making predictions on OOD samples. We hope our preliminary study accelerates the data-centric analysis on various machine learning phenomena.

1 Introduction

Detecting out-of-distribution (OOD) has become one of the key bottlenecks in building reliable open-world systems [3], leading to new SOTA approaches meant to mitigate this problem [5, 16, 18, 29, 20, 23]. Previous works have explored the discrepancy between OOD and in-distribution test performance from the model and algorithm perspective. Le Lan & Dinh (2020) [17] observed the limitation of density estimation for anomaly detection. Zhang et al. (2021) [37] suggested that the cause of OOD failure is the combination of the model architecture and maximum likelihood objective. Choi et al. (2018) [6], Just & Ghosal (2019) [14], Fetaya et al. (2020) [9], Kirichenko et al. (2020) [15], Zhang et al. (2020) [36], and Wang et al. (2020) [34] concluded that significant mismatch between the real and estimated distribution gives rise to the performance discrepancy. Nevertheless, what are the fundamental differences between in-distribution and OOD data samples which cause models to be especially brittle to the latter?

Our approach to studying this difference is inspired by emerging studies on machine learning models adopting *shallow heuristics* (i.e., irrelevant statistical patterns found in the majority of training examples), instead of learning the underlying generalizations that they are intended to capture [22, 33, 1, 2]. For example, Beery et al. (2018) [2] demonstrated a network that accurately recognizes cows in a typical context (e.g., pasture) consistently misclassifies cows in a non-typical context (e.g., beach). Similar heuristics also arise in visual question answering systems [1] and researchers proposed graph generative modeling schemes [13] (inspired by graph convolutional networks [30]) to handle the problem implicitly. In this paper, we study this problem within the Natural Language Inference (NLI): the task of determining whether a premise sentence entails (i.e., implies the truth of) a hypothesis sentence [7, 8, 4]. McCoy et al. (2019) [22] exhaustively characterized shallow

*Corresponding author

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

heuristics that commonly appears in benchmark NLI datasets, which we will refer to as *syntactic heuristics*.

To this end, we study the difference between in-distribution and OOD samples which gives rise to models brittleness through two perspectives: (i) *training dynamics* of the model, and (ii) *syntactic heuristics* of the data samples. Specifically, we perform two types of data-dynamic analyses: We first examine the difference between OOD and in-distribution samples distributions in the data cartography space [28] at each training epoch. Cartography space allows us to distinguish which samples are harder to learn (i.e., the model often misclassifies them) based on training dynamics measures. Second, we mark each data sample by their syntactic heuristic, which enables us to identify what shallow characteristics tend to be harder to learn. For instance, we hypothesize whether more word overlap between a premise and a hypothesis text influences model's ability to infer their label correctly.

Our analyses suggested that the syntactic heuristic that the model deems hard-to-learn during training directly contradicts the characteristic of OOD samples that are hard-to-learn during inference. We also found preliminary evidence suggesting the model's tendency to make inferences based on trivial syntactic heuristic is higher in the OOD case. We hope that this study will drive more effort to better understand the difference between in-distribution and OOD samples and, consequently, develop more informed and data-centric OOD detection and generalization methods.

2 Our Hypotheses

This section presents our two initial hypotheses to characterize OOD data samples. Motivated by previous works [11, 12], we perform a comparative analysis between OOD and in-distribution samples over *dynamic information of model training* (e.g., confidence, prediction variance across epoch) per set of train-test epoch. Moreover, to interpret the quantitative difference of these measures, we mark data samples based on their *syntactic heuristics*. We hypothesize that the difference between OOD and in-distribution samples can be characterized in the combination of multiple training dynamics (§2.1) and sample heuristics (§2.2).

2.1 H1: Are OOD and in-distribution samples different with their training dynamics?

Data cartography [28] is a tool to group data samples into three regions: easy-to-learn, hard-to-learn, and ambiguous (see Figure 1), enabled by following two training dynamics as axes:

1. **Confidence**: the mean model probability of the true label (y_i^*) across epochs (equation 1). The term $p_\theta^{(e)}(y_i^*|x_i)$ denotes the model's probability with parameters θ at the end of eth epoch.

2. **Variability**: the spread of $p_\theta^{(e)}(y_i^*|x_i)$ across epochs, and is defined by equation 2.

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_\theta^{(e)}(y_i^*|x_i) \qquad (1) \qquad\qquad \hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_\theta^{(e)}(y_i^*|x_i) - \hat{\mu}_i)^2}{E}} \qquad (2)$$

Intuitively, higher variability implies a high range of probability outputted by the model for the same sample. As seen in Figure 1, the rightmost region (i.e., high variability) is the ambiguous region. The easy-to-learn region is characterized by high confidence and low variability (i.e., correct prediction with high assigned probability across epochs), and hard-to-learn region samples have low confidence and low variability (i.e., incorrect prediction across epochs).

At each epoch E, we record the following four measurements for all train samples $i \in N_{train}$ and test samples $j \in N_{test}$ (N_{train} and N_{test} denote the size of train and test sets respectively). As θ is constant for all samples at each epoch, for conciseness sake, we abbreviate $p_\theta^{(e)}(y_i^*|x_i)$ as $p_i^{(e)}$.

$$\hat{\mu}_i^E = \frac{1}{E}\sum_{e=1}^{E} p_i^{(e)} \qquad (3) \qquad \hat{\sigma}_i^E = \sqrt{\frac{\sum_{e=1}^{E}(p_i^{(e)} - \hat{\mu}_i^E)^2}{E}} \qquad (4) \qquad \forall \quad i \in N_{train}$$

$$\hat{\mu}_j^E = \frac{1}{E}\sum_{e=1}^{E} p_j^{(e)} \quad (5) \qquad \hat{\sigma_j}^E = \sqrt{\frac{\sum_{e=1}^{E}(p_j^{(e)} - \hat{\mu}_j^E)^2}{E}} \quad (6) \qquad \forall \quad j \in N_{test}$$

In each epoch E, we then plot all train samples based on their respective ($\hat{\mu}_i^E$, $\hat{\sigma_i}^E$) resulting in train cartography map C_{train}^E (e.g., figure 1a). Similarly, we also plot all test samples based on their ($\hat{\mu}_j^E$, $\hat{\sigma_j}^E$) resulting in test cartography map C_{test}^E. (e.g., figure 1c)

As models' performance in inferring in-distribution samples far exceeds that of the OOD case, we hypothesize that most OOD samples will stay in the hard-to-learn regions in C_{test}^E for all E . In contrast, most in-distribution samples will eventually move to the easy-to-learn region in C_{test}^E for larger E, which results in a *noticeable difference between the two distributions in the cartography space.* We hope to interpret this difference using syntactic heuristic (§2.2).

2.2 H2: Do models tend to adopt syntactic attributes of OOD samples more readily?

Past works have shown that machine learning models often adopt 'shortcut' data characteristics instead of the generalization that humans would learn and use to perform the same task [22, 33, 1, 2]. More specifically, a model trained on NLI might assign a label of contradiction to any input containing the word "not" [24, 26], as this heuristic often applies in standard NLI training sets. As OOD samples come from a different distribution, they will have less common generalizable characteristics with the training set than the in-distribution samples. Thus, we hypothesize that *models will rely even more upon the shortcut characteristics (i.e., heuristics) when making inferences on OOD samples.*

We focus on lexical overlap heuristic [22], as it is one of the simplest and most common across NLI datasets. This heuristic assumes that a premise *entails* all hypotheses constructed from words in the premise. Below are examples:

Premise	Hypothesis	Label	Type
The judge was paid by the actor.	The actor paid the judge.	Entailed	Support
The actor was paid by the judge.	The actor paid the judge.	Not entailed	Contradict

A sample can either: (i) supporting, (ii) contradicting (i.e., possessing the property of heuristic and labels *entailed* or *not entailed* respectively), or (iii) having no heuristic. We mark each sample's heuristics, as illustrated by the different circle colors of Figure 1.

To quantify the degree of which each sample adopts lexical heuristic, we calculate the percentage of words in the hypothesis (s2) which overlap with premise (s1), which aligns with the lexical heuristic definition. More formally, we measure $m2 = \frac{|s1 \bigcap s2|}{|s2|}$.

To measure the model's tendency to adapt the heuristic, we calculate three sets of correlations: (i) $\rho(m2, \mu_i^E)$ (ii) $\rho(m2, \mu_{j(in-dist)}^E)$ (iii) $\rho(m2, \mu_{j(OOD)}^E)$, and observe the trends across epochs. We calculate these correlations for all samples and also for samples from each class independently. Intuitively, a high heuristic adoption is indicated by high absolute correlation values in the individual class samples (e.g., the model learns that high words overlap means the label is *entailment*, vice versa).

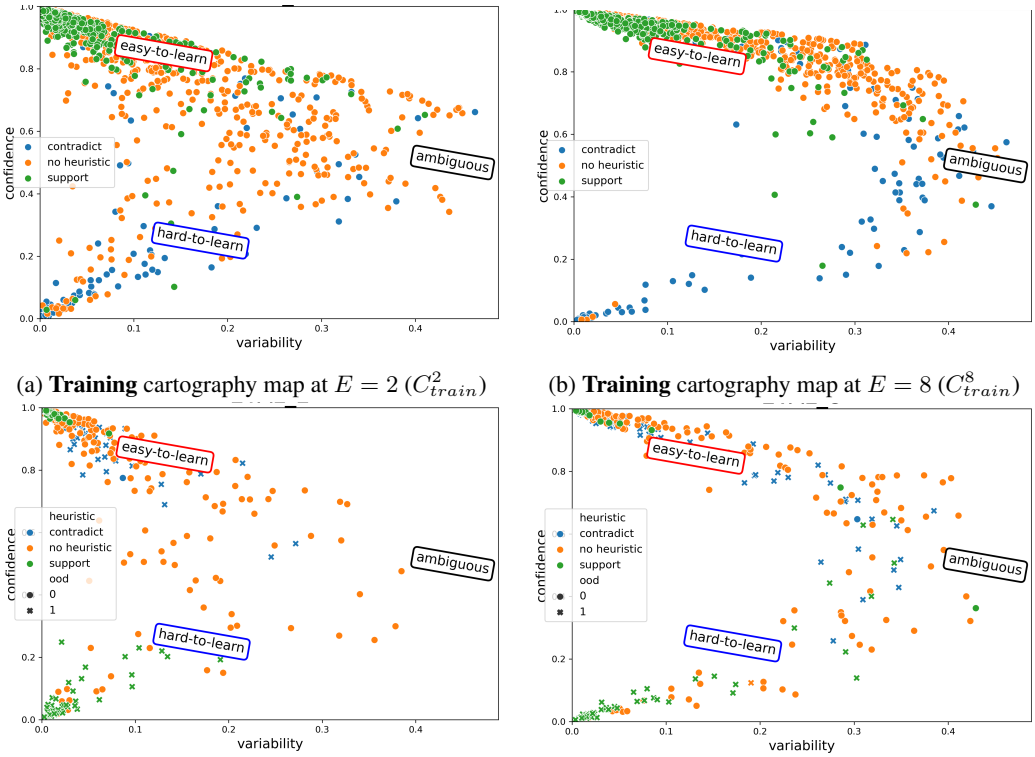
3 Experiments

Setup: In all experiments, we trained Roberta [21] with batch size 20 and learning rate $1.1e^{-5}$, and initialized with the same random seed. All experiments were carried out using PyTorch [25] with one NVidia Tesla K20X GPU. Experiments were carried out using three datasets (MNLI [35], RTE [31], WNLI [19]) which combinations are shown in table 1.

Table 1: Datasets combination in our experiments.

Training	Eval (In-distribution)	Eval (OOD)
MNLI (train)	MNLI (dev matched)	WNLI (train)
MNLI (train)	MNLI (dev matched)	RTE (dev)
RTE (train)	RTE (dev)	WNLI (train)

In our study, we define OOD samples as those with different textual genre and sentence structure from the training. For instance, WNLI were derived by taking sentences from fiction books. Each



(a) **Training** cartography map at $E = 2$ (C^2_{train})

(b) **Training** cartography map at $E = 8$ (C^8_{train})

(c) **Evaluation** cartography map at $E = 2$ (C^2_{test})

(d) **Evaluation** cartography map at $E = 8$ (C^8_{test})

Figure 1: Training and evaluation cartography maps. The different colors illustrate different sample heuristics and the different shapes in figures (c) and (d) illustrate OOD sample ('o' for in-distribution and 'x' for OOD).

hypothesis-premise pair is a sentence and it's copy with the ambiguous pronouns replaced by each possible referent. While RTE and MNLI's hypothesis-premise pairs are different sentences. RTE are texts from Wikipedia and MNLI were crowdsourced from speech, fiction, and government report. Our extensive experiments results can be found in Appendix A.3, Appendix A.1, and Appendix A.2. For conciseness sake, we only present results with MNLI as train and in-distribution data, and WNLI as OOD data.

3.1 OOD vs in-distribution on training dynamics and syntactic heuristic

Result for this section are the cartography maps C^2_{train}, C^8_{train}, C^2_{test}, C^8_{test} in Figure 1. For visualization clarity, the points showed are a sampled fraction from the original set. We observe the following:

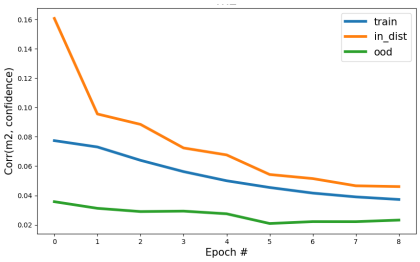
- **Observation:** A contradicting pattern between the trajectories of training and OOD samples. By comparing figures (a) and (b) (C^2_{train} and C^8_{train}), we observed samples that *contradict the heuristic* (blue circles) either stays in the hard-to-learn region or move towards the ambiguous region. On contrary, the figures (c) and (d) (C^2_{test} and C^8_{test}) shows that OOD samples that *support the heuristic* (green 'x's) stay in the hard-to-learn region.
  Conjecture: Samples heuristics (i.e., lexical overlap) that the model deems hard/easy to learn during training *completely flips in the OOD case*.
- **Observation:** At the end of epoch 8, we observe way less color mix at the ambiguous region of C^8_{test} (blue and green 'x's at the ambiguous region of map (d)). While in C^8_{test} (map (b)), this mix is very apparent in the easy-to learn region (blue and green circles).
  Conjecture: a more generalizable knowledge is learned during training (indicated by the samples mix). Although, *this generalizable knowledge is slower to be adopted by the model in inferring OOD samples*.

3.2 OOD vs in-distribution on tendency to adopt syntactic heuristics

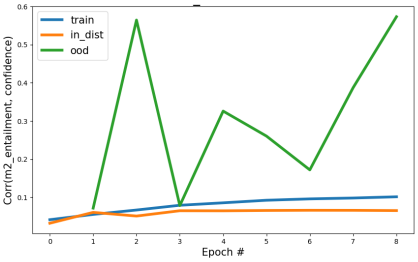
For the second hypothesis, we find the following observations (refer to figure 2):

- **Observation:** Correlation values ($\rho(m2, \mu_i^E)$, $\rho(m2, \mu_{j(in-dist)}^E)$, $\rho(m2, \mu_{j(OOD)}^E)$) are relatively low in the plot for all samples (a). This applies to all cases (train, in-distribution, and OOD test). However, *an obvious divergence is observed in the* $\rho(m2, \mu_{j(OOD)}^E)$ *plot for the entailment samples (b).*
  Conjecture: This observation confirms our hypothesis in 2.2 that models are *more prone to adopt syntactic heuristics when making inference on OOD samples*.

- **Observation:** All samples' trends adhere to the entailment trends in the train and in-distribution test case (orange and blue lines in figures (a) and (b)) but contradict the OOD case (green lines in figures (a) and (b)). In the OOD case, we can see that the negative and non-zero slopes are only found in all samples trend, but the correlation keeps increasing (with a huge difference in value) in the entailment samples trend.
  Conjecture: This might indicate that knowledge used by the model to infer class labels is *consistent for all classes in train and in-distribution but conflicts in the OOD case*.



(a) Correlation in **all** samples



(b) Correlation in **entailment** samples

Figure 2: Hypothesis 2.2: $\rho(m2, \mu_i^E)$, $\rho(m2, \mu_{j(in-dist)}^E)$, $\rho(m2, \mu_{j(OOD)}^E)$ on *all* samples (a) and *entailment* samples(b)

More supplementary plots of this analysis (more experiments on different datasets and the trend plots for **non-entailment** samples) can be found in appendix A.4.

4 Limitations and Future Directions

This data-centric approach to understanding the difference between OOD and in-distribution samples offers an intuition of failure mode in OOD. However, the results and approach presented in this paper have some limitations.

Although all NLI sets in our experiments came from different topic domains and had distinct characteristics, a more careful selection is needed to ensure maximum separation of in-distribution vs. OOD data distribution. Zhang et al. (2021) [37] highlights the problem with experiments setup where OOD regions can lie in support of the data distribution: high logit values assigned to certain OOD samples due to model estimation error. By ensuring this criterion is met, we can achieve a more accurate confidence score, which in turn will improve our results in section 3.1.

More controlled experiment is needed for results in section 3.2 to ensure that the correlation between confidence and m2 is not spurious (i.e., there exists no unobserved confounding variable). One way to remedy this is by running the same measurements on random subsets and see if the correlation pattern still holds (i.e., randomized controlled experiments).

Future work may include remedies for the two mentioned limitations. Furthermore, investigating the root cause of contradicting syntactic heuristics of OOD and in-distribution samples that the model fails to predict correctly (section 3.1) is also an exciting direction. It will also be exciting to see the effect of correcting for distribution mismatches [27, 10, 32] to the observations presented in this paper. To further corroborate the analysis and conjecture presented in this paper, the two possible directions can be: (i) adding more syntactic heuristics for analysis (e.g., coreference, negation), and (ii) setting up similar experiments on vision and other domains datasets. Finally, we hope that the insights provided by this data-centric view can motivate more informed development of OOD detection and generalization methods.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 472–489, Cham, 2018. Springer International Publishing.

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, September 2015. Association for Computational Linguistics.

[5] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

[6] Hyun-Jae Choi, Eric Jang, and Alexander Amir Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv: Machine Learning*, 2018.

[7] Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45, 2003.

[8] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[9] Ethan Fetaya, Jörn-Henrik Jacobsen, Will Grathwohl, and Richard Zemel. Understanding the limitations of conditional generative models. *arXiv preprint arXiv:1906.01171*, 2019.

[10] Alexander Fuchs, Christian Knoll, and Franz Pernkopf. Distribution mismatch correction for improved robustness in deep neural networks. *arXiv preprint arXiv:2110.01955*, 2021.

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

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

[13] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 199–208, 2021.

[14] John Just and Sambuddha Ghosal. Deep generative models strike back! improving understanding and evaluation in light of unmet expectations for ood data. *arXiv preprint arXiv:1911.04699*, 2019.

[15] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data, 2020.

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc.

[17] Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *arXiv preprint arXiv:2012.03808*, 2020.

[18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[19] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[20] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020.

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

[22] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[23] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020.

[24] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[26] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. 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 1975–1985, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[27] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 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, November 2020. Association for Computational Linguistics.

[29] Kshitij Tayal, Rahul Ghosh, and Vipin Kumar. Model-agnostic methods for text classification with inherent noise. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 202–213, 2020.

[30] Kshitij Tayal, Nikhil Rao, Saurabh Agarwal, Xiaowei Jia, Karthik Subbian, and Vipin Kumar. Regularized graph convolutional networks for short text classification. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 236–242, 2020.

[31] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

[32] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[33] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. Visual concepts and compositional voting. *arXiv preprint arXiv:1711.04451*, 2017.

[34] Ziyu Wang, Bin Dai, David Wipf, and Jun Zhu. Further analysis of outlier detection with deep generative models, 2020.

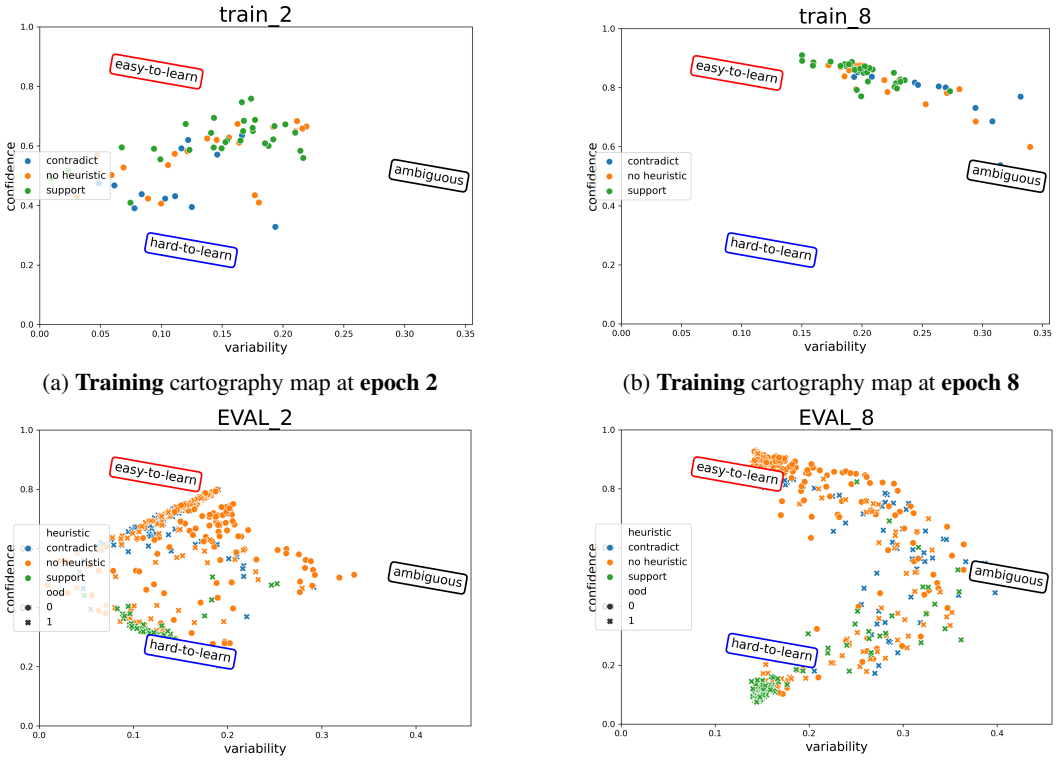
[35] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[36] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 102–117, Cham, 2020. Springer International Publishing.

[37] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021.

A Appendix

A.1 Supplementary plot: OOD vs in-distribution on training dynamics information (Training and in-dis: RTE; OOD: WNLI)



(a) **Training** cartography map at **epoch 2**

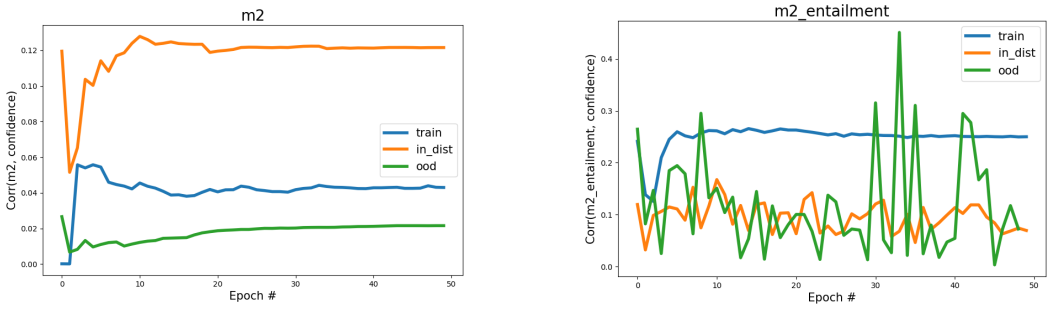
(b) **Training** cartography map at **epoch 8**

(c) **Evaluation** cartography map at **epoch 2**

(d) **Evaluation** cartography map at **epoch 8**

Figure 3: Training cartography maps (training set: RTE). The number of heuristics related samples in RTE is small.

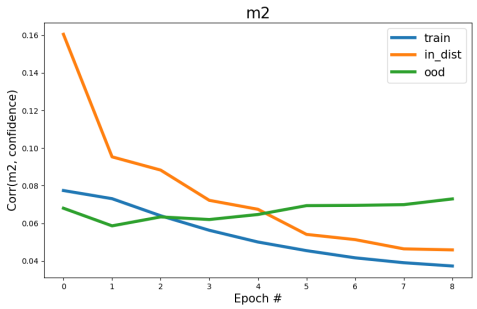
A.2 Supplementary plot: OOD vs in-distribution on syntactic characteristics (entailment)



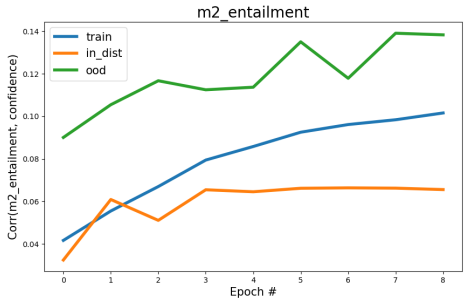
(a) Correlation between **m2** and **all** samples $\hat{\mu}_i$

(b) Correlation between **m2** and **entailment** samples $\hat{\mu}_i$

Figure 4: Results for hypothesis 2.2. Training and in-distribution test samples are RTE, and OOD samples are WNLI.



(a) Correlation between **m2** and **all** samples $\hat{\mu}_i$

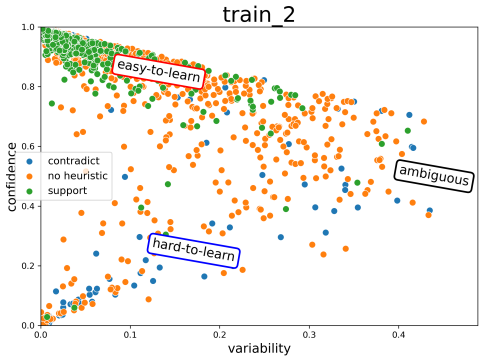


(b) Correlation between **m2** and **entailment** samples $\hat{\mu}_i$

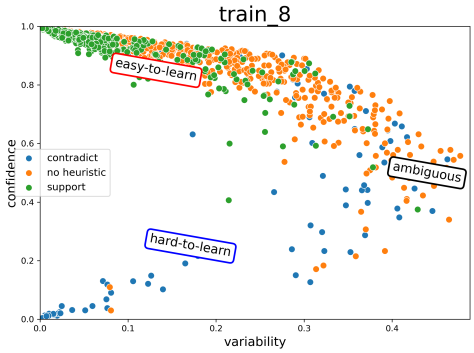
Figure 5: Results for hypothesis 2.2. Training and in-distribution test samples are MNLI, and OOD samples are RTE.

Results presented are at the end of epoch 8 for MNLI training and the end of epoch 50 for RTE training. This is based on the epoch in which the training error has converged (around 0.02).

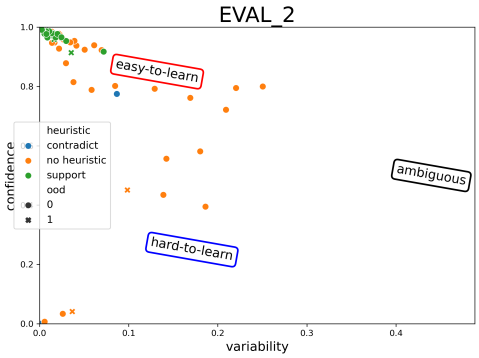
A.3 Supplementary plot: OOD vs in-distribution on training dynamics information (Training and in-dis: MNLI; OOD: RTE)



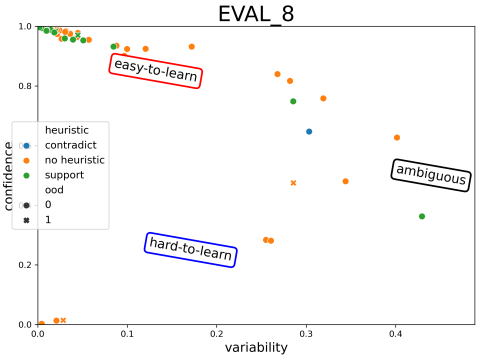
(a) **Training** cartography map at **epoch 2**



(b) **Training** cartography map at **epoch 8**



(c) **Evaluation** cartography map at **epoch 2**

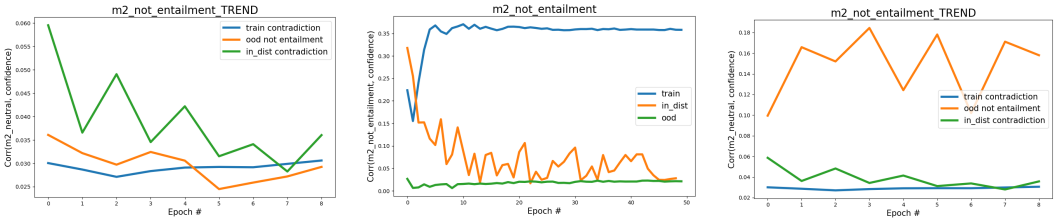


(d) **Evaluation** cartography map at **epoch 8**

Figure 6: Training and evaluation cartography maps (train: MNLI, evaluation: RTE). The number of heuristics related samples in RTE is small.

A.4 Supplementary plot: OOD vs in-distribution on syntactic characteristics (non-entailment)

This section shows plots for correlation between confidence scores ($\hat{\mu}_i$) of **non entailment** samples and m2

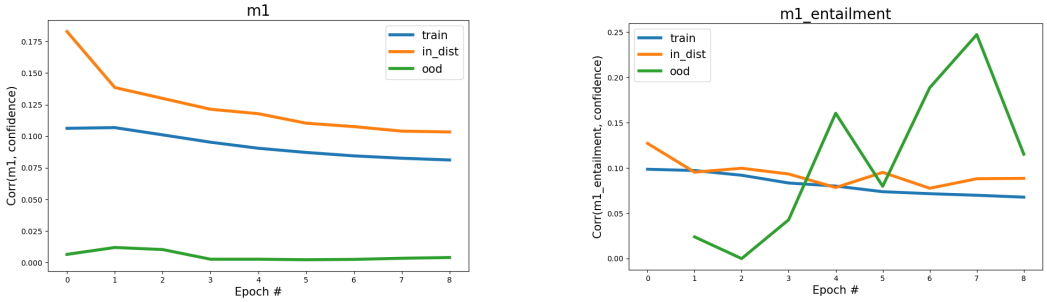


(a) Train & in-distribution: MNLI, OOD: WNLI (b) Train & in-distribution: RTE, OOD: WNLI (c) Train & in-distribution: MNLI, OOD: RTE

Figure 7: Supplementary results for 3.2. Correlation between $\hat{\mu}_i$ of **non-entailment** samples and m2

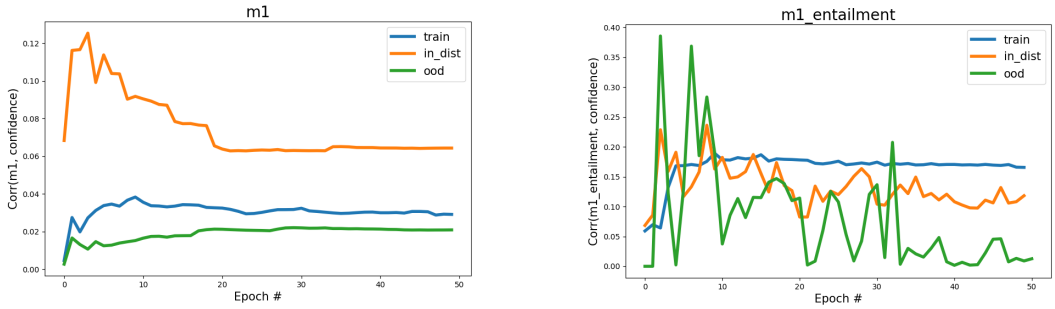
A.5 Supplementary material: Extra lexical overlap measure

We also added another measure to quantify tendency to adopt lexical overlap heuristic. We calculated $m1 = \frac{|s1 \bigcap s2|}{|s1|}$. Essentially, this measures how much percentage of words found in the premise ($s1$) can also be found in the hypothesis ($s2$).



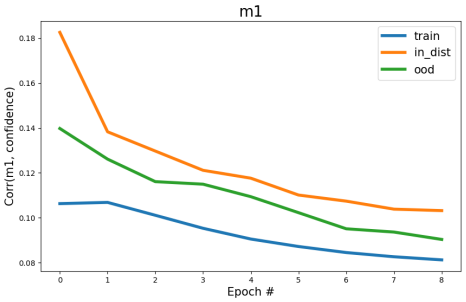
(a) Correlation between **m2** and **all** samples $\hat{\mu}_i$ (b) Correlation between **m2** and **entailment** samples $\hat{\mu}_i$

Figure 8: Results for hypothesis 2.2. Training and in-distribution test samples are MNLI, and OOD samples are WNLI.

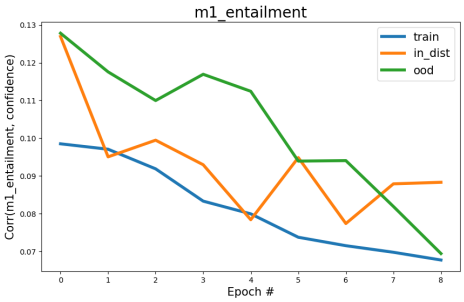


(a) Correlation between **m2** and **all** samples $\hat{\mu}_i$ (b) Correlation between **m2** and **entailment** samples $\hat{\mu}_i$

Figure 9: Results for hypothesis 2.2. Training and in-distribution test samples are RTE, and OOD samples are WNLI.



(a) Correlation between **m2** and **all** samples $\hat{\mu}_i$



(b) Correlation between **m2** and **entailment** samples $\hat{\mu}_i$

Figure 10: Results for hypothesis 2.2. Training and in-distribution test samples are MNLI, and OOD samples are RTE.