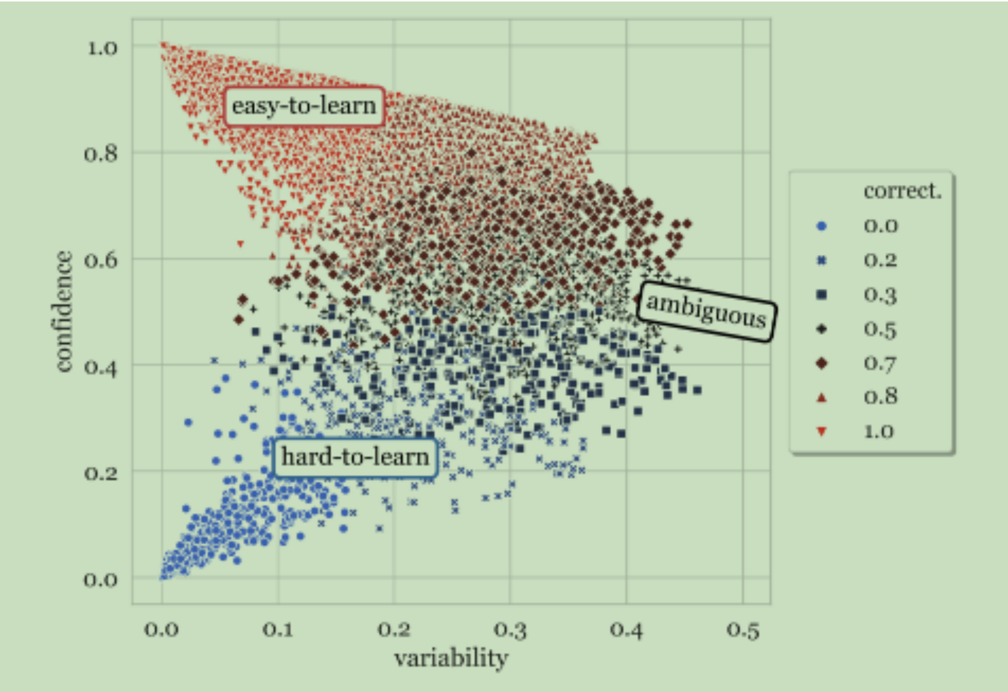
Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics

Abstract

- This yields two intuitive mea sures for each example—the model's confidence in the true class, and the variability of this confifidence across epochs—obtained in a single run of training
  - 观察模型的置信度和偏差的分布
- The common belief is that the more abundant the labeled data, the higher the likelihood of learning diverse phe nomena, which in turn leads to models that gener alize well
- Training on ambiguous instances pro motes generalization to OOD test sets, with little or no effect on in-distribution (ID) performance
- Easy case 可以加快模型收敛
- 清洗错误数据，扩充易混淆数据，少量容易数据，训练的又快，泛化能力又好
- 通过对数据的置信度和偏差的分布观察
  - 

  - 置信度高 且 偏差小的数据，是 easy case，对模型的优化很关键
  - 置信度一般 且 偏差大的数据，是 ambiguous case，对模型的泛化能力很关键
  - 置信度低 且 偏差小的数据，是 error case，一般是错误数据，可以用于清洗数据集
- 大量的、多样性强的数据，对模型的泛化能力更关键
- 先从简单的样本学起，模型收敛的更快
- 易混淆样本占比在 25% 左右，泛化效果才会明显，低于 17% 没啥用，大于 25% 会有反向效果
- 低置信度的样本中，可能包括错误标签
- 作者用的置信度和方差，是在多个 epoch 中的均值，而不是最后一波预测的，这样可以找到再训练中的 esay case 和 hard case，结果更平滑、也更置信

Data Selection using Data Maps

- 需要分析不同区域的数据对于模型的学习和泛化能力的区分
-

		WinoG. Val. (ID)	*WSC* (OOD)
	100% train	$79.7_{0.2}$	$86.0_{0.1}$
33% train	random	$73.3_{1.3}$	$85.6_{0.4}$
	high-**correctness**	$70.8_{0.6}$	$84.1_{0.4}$
	high-**confidence**	$69.4_{0.5}$	$83.9_{0.5}$
	low-**variability**	$70.1_{1.0}$	$83.7_{1.4}$
	forgetting	$75.5_{1.3}$	$84.8_{0.7}$
	AL-uncertainty	$75.7_{0.8}$	$85.7_{0.8}$
	AL-greedyK	$74.2_{0.4}$	$86.5_{0.5}$
	AFLite	$76.8_{0.8}$	$86.6_{0.6}$
	low-**correctness**	$78.2_{0.6}$	$86.3_{0.6}$
	hard-to-learn	$77.9_{1.3}$	$87.2_{0.7}$
	ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Table 2: ID and OOD accuracies for RoBERTa-large models trained on different selections of *WinoGrande*. Reported values are averaged over 3 random seeds, with s.d. reported as a subscript. Selection of 33% training instances with highest **variability** (*ambiguous*) achieves the best OOD performance, outperforming all other baselines from this work, as well as prior work.

		SNLI						*MultiNLI*						
		ID	*NLI Diagnostics* (OOD)					ID (Val.)		*NLI Diagnostics* (OOD)				
		Test	Lex.	PAS	LS	Kno.	All	Mat.	MisM.	Lex.	PAS	LS	Kno.	All
	100% train	92.0	54.6	67.9	62.7	52.1	61.8	**90.2**	**90.1**	59.9	68.4	67.3	57.8	65.0
33% train	*random*	91.3	53.0	66.8	59.7	50.7	60.4	89.8	89.2	59.3	69.6	66.5	56.3	64.6
33% train	*hard-to-learn*	91.8	55.2	**69.1**	63.2	51.7	62.0	89.5	89.7	59.3	68.9	**69.5**	58.8	65.3
33% train	*ambiguous*	**92.2**	**58.5**	67.9	**64.1**	**54.2**	**63.5**	90.1	89.3	**63.5**	**71.0**	68.9	**59.2**	**66.9**

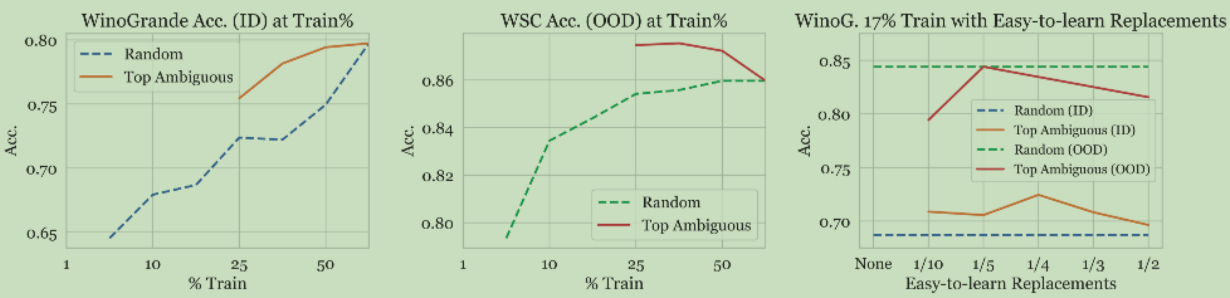
  Table 3: ID and OOD accuracies for RoBERTa-large models trained on different selections of *SNLI* and *MultiNLI*; we report the best performance over 3 random seeds (see Appendix §B for *SNLI* validation results). *ambiguous* and *hard-to-learn* subsets of data promote OOD generalization, at minimal degradation of ID performance. OOD performance improves across all fine-grained linguistic categories in the *NLI Diagnostics* set.
- 

  Figure 3: ID (left) and OOD (centre) *WinoGrande* performance with increasing % of *ambiguous* (and randomly-sampled) training data. RoBERTa-large optimization fails when trained on small amounts (< 25%) of the most *ambiguous* data (results correspond to majority baseline performance and are not shown here, for better visibility). (Right) Replacing small amounts of *ambiguous* examples from the 17% subset with *easy-to-learn* examples results in successful optimization and ID improvements, at the cost of decreased OOD accuracy. All reported performances are averaged over 3 random seeds.
- hard case 部分，可能包括了错误数据