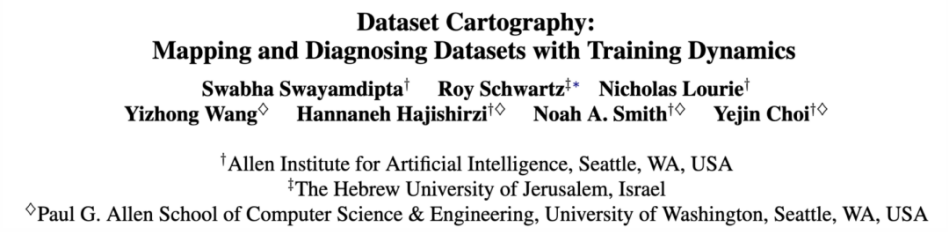
数据地图---使用Training Dynamics来映射和诊断数据集

原创 郭必扬 SimpleAI 2022-07-11 17:22 发表于上海

数据地图---使用Training Dynamics来映射和诊断数据集

最近看到一篇很有趣的文章，发表于EMNLP-20，作者团队主要来自AllenAI：

Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics



Dataset Cartography:
Mapping and Diagnosing Datasets with Training Dynamics

Swabha Swayamdipta[†] Roy Schwartz[‡*] Nicholas Lourie[†]
Yizhong Wang[◇] Hannaneh Hajishirzi[†◇] Noah A. Smith[†◇] Yejin Choi[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[‡]The Hebrew University of Jerusalem, Israel
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

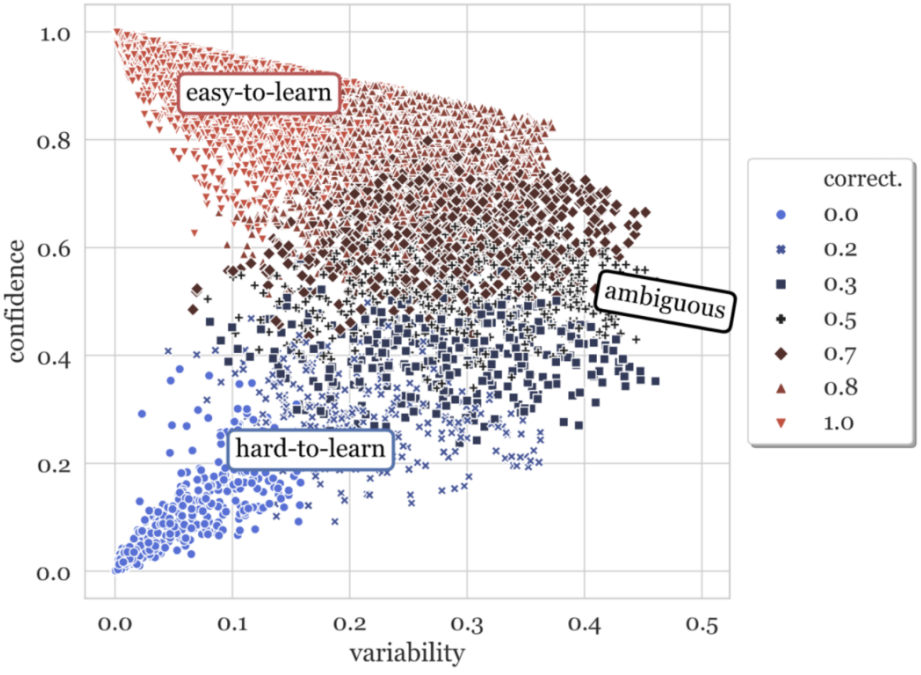
我们以往的关注点主要在模型身上，这篇文章则是关注于我们的训练数据集，希望通过模型训练过程中的一些动态指标——training dynamics，来发掘数据集的一些性质，比如不同样本的难易程度，从而帮助我们更好地训练模型。这其实是Data-centric方向中的data selection要考虑的主要问题之一。

曾经我介绍过另一篇分析训练过程中的example forgetting现象的文章（深度学习中的样本遗忘问题 (ICLR-2019)），这篇文章则是在此基础上更进一步，用一种更精细化的方式，来可视化我们的数据集。

论文的核心方法，用一句话就可以介绍完毕：

> 假设我们训练一个分类模型N个epoch，针对每一个sample，在每个epoch结束后，我们都记录该sample在正确类别上的概率。然后在训练结束后，我们对这N个概率，**我们计算概率的均值和标准差，分别记为confidence和variability，构成该sample的坐标，这样就可以绘制数据地图（dataset cartography）**。

下面是使用SNLI数据集绘制的数据地图：



上图大致可以分为三个区域：

- **easy-to-learn**：是confidence较高，但是variability较低的区域
- **hard-to-learn**：是confidence较低，variability也较低的区域
- **ambiguous**：是variability较高的区域

从名字就可以看出这三个区域的样本，拥有不同的性质。

接下来作者做了一个实验，只使用某一个区域的样本进行训练，看看分别有什么样的效果：

		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	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

上面这个表中，作者只选取了1/3的样本，来跟全量样本的训练进行对比。high-confidence就是指easy-to-learn的样本。可以看出：

- 只使用easy的样本，效果会很差，比随机选1/3的结果都差；
- 只使用hard的样本，效果不错，在OOD上甚至可以超过100%训练样本
- 只使用ambiguous样本，在所有subset中效果最好

在其他数据集上，也有类似的现象：

		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
	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
	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**

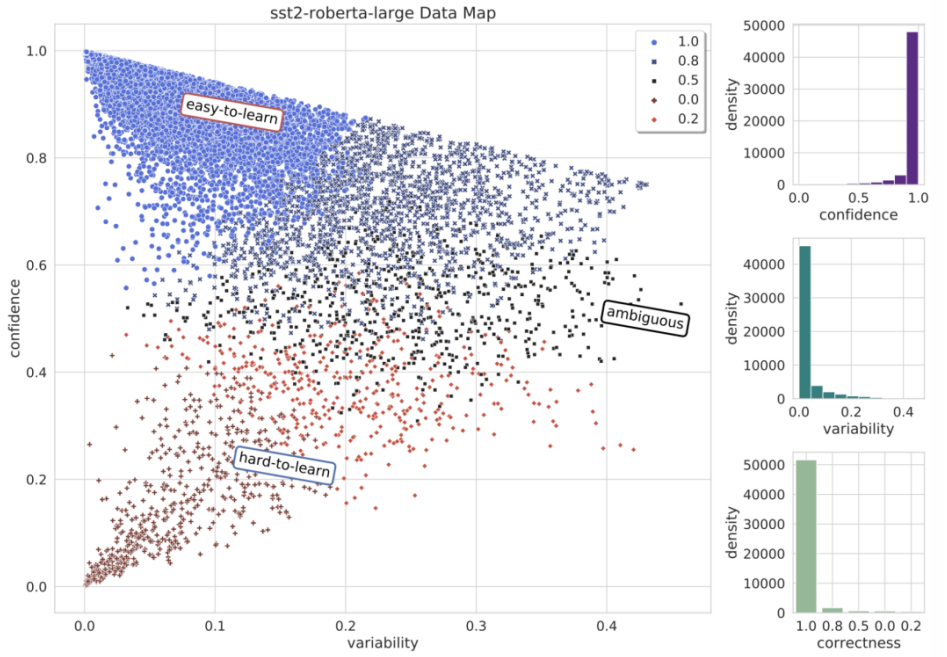
作者进一步做了一些实验，来探究三个区域样本的功能，发现：

- easy样本，虽然对模型性能的贡献不大，但是如果完全不使用的话，模型的收敛会很困难
- ambiguous的贡献基本上是最大的
- hard样本贡献也很大，但是里面可能包含很多noise，如果数据错标的话，基本都出现在hard区域

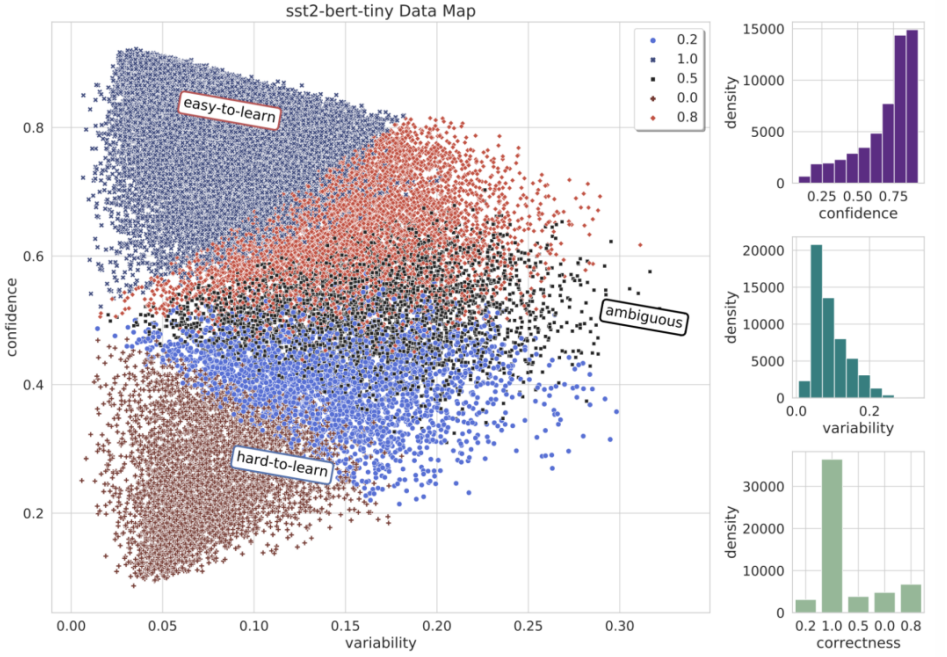
以上就差不多是论文的内容了，其实很简单，但是这样的一个数据地图，其实可以帮助我们进一步观察数据集的特点，帮助我们从data-centric的角度去做出改进。

笔者自己也跑了一下在SST2数据集上的数据地图，分别使用一个大模型和一个小模型，发现差异明显：

下图是使用RoBERTa-large的效果：



下图则是使用BERT-tiny的效果：



还是挺有意思的，通过这些差异，也许我们可以进一步地发现数据集中的一些特点。

原作者的GitHub：*https://github.com/allenai/cartography*

然而这个repo好久没有维护了，很难直接运行，所以我使用最新版的transformers库复现了一下，两行命令即可绘制上述数据地图：

https://github.com/beyondguo/TrainingDynamics

欢迎大家 star 来跑跑看。

写作不易

如果觉得有所收获的话

大家就点一个**赞**吧 :)

SimpleAI
追求用简单、有趣的方式来分享AI知识。
89篇原创内容

公众号

2022年的第10/52篇原创笔记
和我一起挖掘有趣的AI研究吧！

近期笔记推荐：

盘点Controllable Text Generation(CTG)的进展
文本检索、开放域问答与Dense Passage Retrieval
用Annoy和ThreadPool把相似度计算加速360倍
劫富济贫：对长尾数据进行特征空间增强
通俗科普文：贝叶斯优化与SMBO、高斯过程回归、TPE
深度学习中的样本遗忘问题 (ICLR-2019)