Sampling Bias in Deep Active Classification

An Empirical Study

Ameya Prabhu*#[1]; Charles Dognin*[2]; Maneesh Singh[2]

[1]University of Oxford, [2]Verisk|AI

* Equal Contribution # Work done at Verisk|AI

Motivation

Goal: Efficient data sampling for training large DNN classifier models.

- Literature finds that uncertainty sampling is biased
- Selects redundant samples from a region in f
- Does not scale with batched sampling

Recent works propose approaches to alleviate these:

- Diversity sampling
- Bayesian sampling and Ensembles

Use Cases

- Annotate powerful small labeled datasets using active learning with fast compact models
- Perform fast, compute-efficient training of large transformer models at minimal accuracy loss

Setup

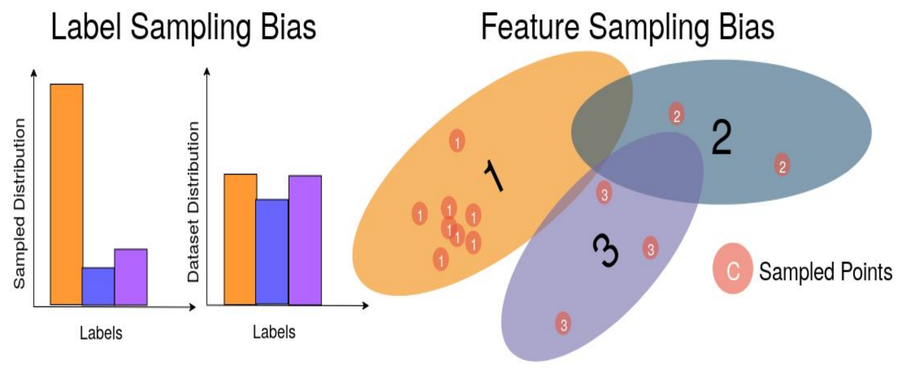
- Setup Details:
  - 8 large text classification datasets (Size 120K-3.6M)
  - Two models: TFIDF-MNB & FastText.zip representating traditional & deep models
  - No sources of randomness: deterministic setup
- Hypothesis tested:
  - Label and distributional sampling bias
  - Algorithmic factors: Initial set selection, query size and query strategy with two trained models and four acquisition functions
- Large scale empirical study: Combinatorial explosion of factors, large scale study necessary to isolate effects. Hence, we run ~2300 experiments.
- Uniqueness of Study
  - Our datasets 2 orders of magnitude larger
  - Query size often ~dataset size of past works
  - Extensive benchmarking (20x experiments)

We present trends consistent across all 8 datasets and robust across various isolation settings

Popular active learning (AL) hypotheses tested for deep models

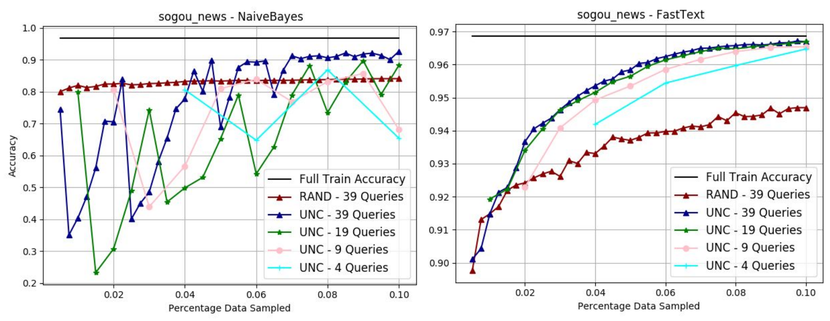
✗ Uncertainty based AL is label-biased & feature-biased

Dsets	Limit	FTZ ($\cap Q$)	MNB ($\cap Q$)	FTZ ($\cap S$)	MNB ($\cap S$)
SGN	1.61	1.56 ± 0.03	1.15 ± 0.32	1.59 ± 0.01	1.57 ± 0.01
DBP	2.64	2.50 ± 0.02	2.27 ± 0.11	2.51 ± 0.0	2.58 ± 0.01
YHA	2.30	2.25 ± 0.01	2.22 ± 0.02	2.25 ± 0.0	2.28 ± 0.0
YRP	0.69	0.69 ± 0.0	0.56 ± 0.13	0.69 ± 0.0	0.69 ± 0.01
YRF	1.61	1.56 ± 0.02	1.42 ± 0.21	1.56 ± 0.0	1.57 ± 0.01
AGN	1.39	1.33 ± 0.04	1.13 ± 0.17	1.33 ± 0.0	1.35 ± 0.01
AMZP	0.69	0.69 ± 0.0	0.69 ± 0.0	0.69 ± 0.0	0.69 ± 0.0
AMZF	1.61	1.58 ± 0.02	1.6 ± 0.01	1.59 ± 0.0	1.61 ± 0.0



✗ Uncertainty based AL irreversibly degrades with ⬆ in query size

Dsets	Chance	FTZ 9 ∩ 19 ∩ 39	FTZ 39 ∩ 39 ∩ 39	MNB 9 ∩ 19 ∩ 39	MNB 39 ∩ 39 ∩ 39
SGN	0.83 ±0.0	77.0 ± 0.5	77.9 ± 0.2	31.9 ± 0.0	55.5 ± 0.0
DBP	0.9 ±0.0	80.0 ± 0.1	79.6 ± 0.2	82.3 ± 0.0	79.7 ± 0.0
YHA	3.7 ±0.0	68.3 ± 0.1	69.0 ± 0.0	92.1 ± 0.0	89.5 ± 0.0
YRP	0.9 ±0.0	46.0 ± 0.9	42.7 ± 1.0	10.8 ± 0.0	16.0 ± 0.0
YRF	3.6 ±0.0	68.4 ± 0.2	67.6 ± 0.1	14.2 ± 0.0	13.6 ± 0.0
AGN	3.7 ±0.0	70.3 ± 0.2	68.7 ± 0.1	81.6 ± 0.0	79.8 ± 0.0
AMZP	0.9 ±0.0	45.8 ± 0.1	48.2 ± 0.2	11.5 ± 0.0	15.0 ± 0.0
AMZF	3.6 ±0.0	55.2 ± 0.4	57.0 ± 0.2	28.4 ± 0.0	57.8 ± 0.0

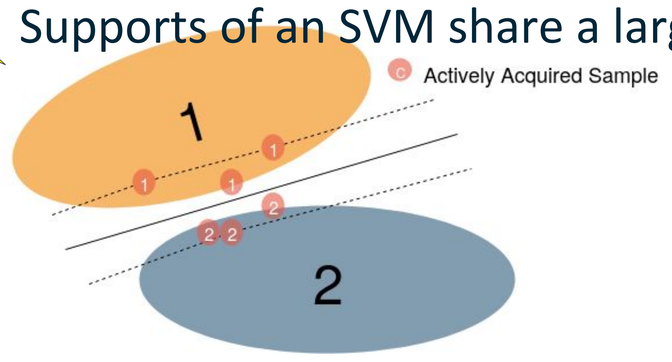


✗ Ensembles help improve quality of sampling in active classification

Dsets	Chance	FTZ Ent-Ent	FTZ Ent-LC	FTZ Ent-DelEnt	FTZ DelEnt-DelLC	FTZ DelEnt-DelEnt
SGN	9.4 ± 0.0	84.6 ± 0.2	83.1 ± 0.3	81.7 ± 0.1	82.6 ± 0.1	84.2 ± 0.1
DBP	9.3 ± 0.0	85.7 ± 0.2	85.5 ± 0.3	83.3 ± 0.1	83.0 ± 0.4	83.2 ± 0.2
YHA	19.0 ± 0.0	79.0 ± 0.0	71.6 ± 0.2	76.3 ± 0.1	69.6 ± 0.7	75.6 ± 3.9
YRP	9.3 ± 0.0	58.4 ± 0.6	59.0 ± 0.3	59.0 ± 0.6	61.6 ± 0.7	62.1 ± 0.1
YRF	19.0 ± 0.0	77.8 ± 0.2	66.6 ± 0.3	75.8 ± 0.1	65.4 ± 0.3	80.1 ± 0.2
AGN	19.1 ± 0.0	78.3 ± 0.1	77.3 ± 0.1	77.1 ± 0.3	78.2 ± 0.4	79.0 ± 0.3
AMZP	9.5 ± 0.0	63.5 ± 0.2	63.5 ± 0.3	66.1 ± 0.4	70.0 ± 0.1	70.0 ± 0.1
AMZF	19.0 ± 0.0	70.3 ± 0.1	64.3 ± 0.2	69.6 ± 0.1	65.6 ± 0.2	72.6 ± 0.2

Dsets	Chance	FTZ-FTZ Ent	FTZ-5F TZ Ent	5FTZ-5FTZ Ent-LC	5FTZ-5FTZ Ent-Ent
SGN	9.4 ± 0.0	84.6 ± 0.2	86.3 ± 0.2	85.4 ± 0.4	85.8 ± 0.0
DBP	9.3 ± 0.0	85.7 ± 0.2	86.6 ± 0.3	86.78 ± 0.1	87.8 ± 0.2
YRP	9.3 ± 0.0	58.4 ± 0.6	58.1 ± 0.7	58.3 ± 0.3	58.2 ± 0.2
YRF	19.0 ± 0.0	77.8 ± 0.2	79.0 ± 0.3	68.5 ± 1.1	77.6 ± 0.3
AGN	19.1 ± 0.0	78.3 ± 0.1	79.0 ± 0.2	79.1 ± 0.2	77.9 ± 0.2

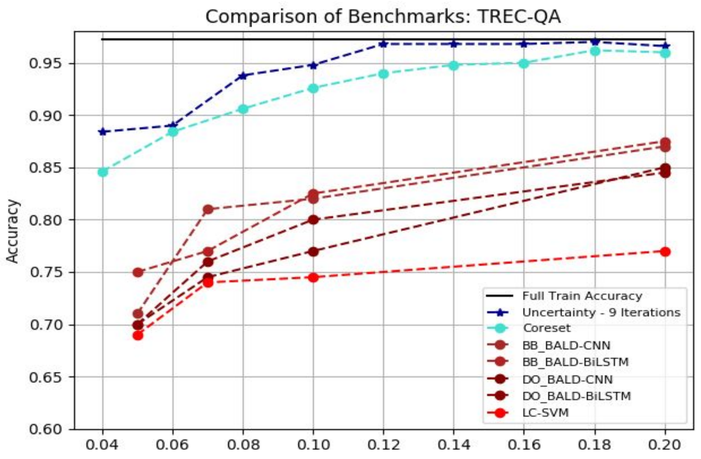
Supports of an SVM share a large overlap with our acquired points

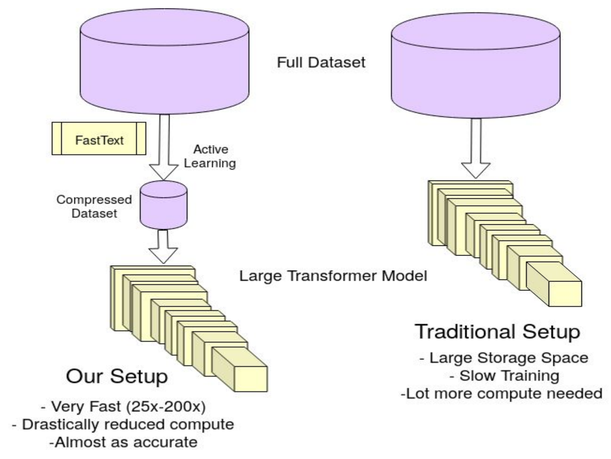


Dsets	Common%	Chance%	$\#SV$
SGN	71.3 ± 0.5	9.3 ± 0.5	13184
DBP	86.3 ± 0.5	9.7 ± 0.5	1479
YRP	57.3 ± 0.5	9.7 ± 0.5	31750
AGN	45.0 ± 0.8	21.0 ± 1.6	1032

Table 3: Proportion of Support Vectors intersecting with our actively selected set calculated by $\frac{|\mathcal{S}_{SV} \cap \hat{\mathcal{S}}_b|}{|\mathcal{S}_{SV}|}$.

Applications





Model	AGN	DBP	SGN	YRF	YRP	YHA	AMZP	AMZF
VDCNN (Conneau et al., 2017)	91.3	98.7	96.8	64.7	95.7	73.4	95.7	63.0
DPCNN (Johnson and Zhang, 2017)	93.1	**99.1**	**98.1**	**69.4**	**97.3**	**76.1**	**96.7**	**65.2**
WC-Reg (Qiao et al., 2018)	92.8	98.9	**97.6**	64.9	96.4	73.7	95.1	60.9
DC+MFA (Wang et al., 2018)	93.6	**99.2**	-	66.0	96.5	-	-	63.0
DRNN (Wang, 2018)	**94.5**	**99.2**	-	**69.1**	**97.3**	70.3	**96.5**	**64.4**
ULMFiT (Howard and Ruder, 2018)	**95.0**	**99.2**	-	**70.0**	**97.8**	-	-	-
EXAM (Du et al., 2019)	93.0	99.0	-	-	-	74.8	95.5	61.9
Ours: ULMFiT (Small data)	93.7 (20)	**99.2 (10)**	97.0 (10)	67.6 (20)	**97.1 (10)**	74.3 (20)	**96.1 (10)**	**64.1 (20)**
Ours: ULMFiT (Tiny data)	91.7 (8)	98.6 (2.3)	**97.4 (6.3)**	66.3 (8)	96.7 (4)	73.3 (8)	95.8 (4)	62.9 (8)

Take Aways

Summary

Uncertainty based AL with deep models like Fasttext.zip show:

- Negligible class bias
- No adverse feature bias (favorable)
- Scales with query size (no degradation)

Surprising Discoveries

- Ensembling does not improve sampling
- Supports of a SVM have large overlap with our acquired samples

Uses

- Generates compact surrogate datasets
  - Speedup large DNN training by 25-200x
- State-of-the-art in deep active text classification
  - Outperforms prev. best by 4x less data

Contact

Ameya Prabhu
ameya.prabhu@mailfence.com

Charles Dognin
charles.dognin@verisk.com

Code

https://github.com/drimpossible/Sampling-Bias-Active-Learning

References

1. Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In EMNLP 2018
2. Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In ICLR 2018
3. Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. ArXiv preprint arxiv:1907.06347v1