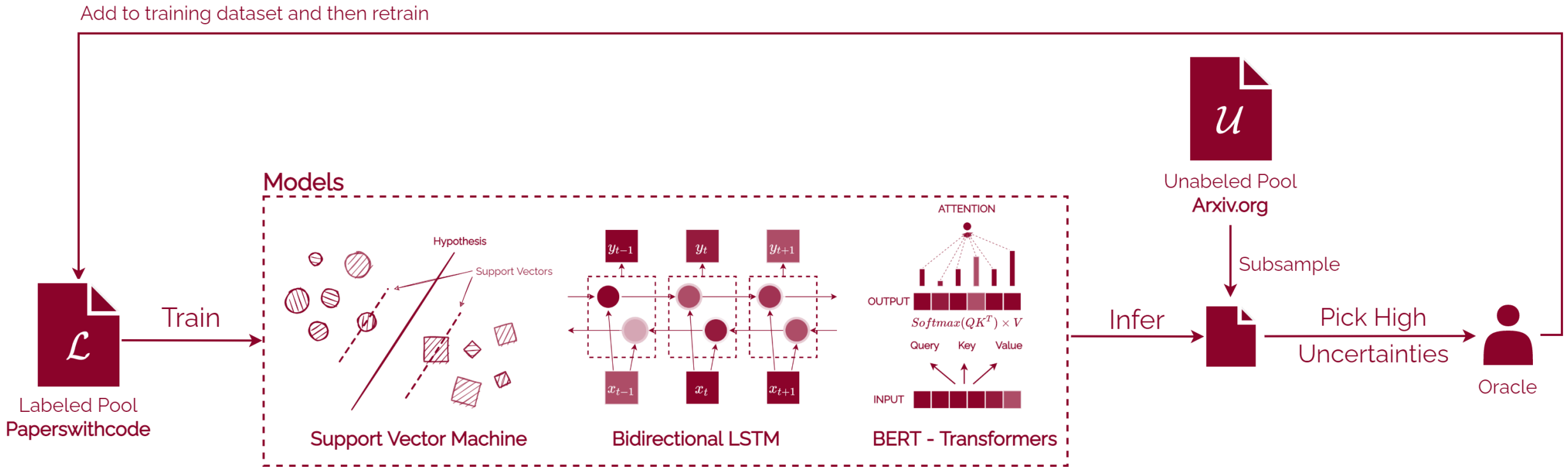
Paper Topic Classification with Active Learning

Daehyun Cho
Cognitive System Lab
A.I. Dept, Korea Univ.

Index

Whole Pipeline Blueprint

※ Some conditions may change during the project



Multi-label Classification Problem with Paper Title+Abstract

Data

- paperswithcode: contains 12k papers from A.I. Field
- arxiv: contains 1.7M papers from many fields – will use A.I. related only

Task

- A.I. Field Multi-label Classification
- Active Learning with Labeled Pool (paperswithcode) and Unlabeled Pool (arxiv)

Details

- Uncertainty-based sampling methods
  - Classic Active Learning / BALD / BatchBALD
- Uncertainty-estimation methods
  - Monte Carlo Dropout / Full Ensemble
- and compare with Naïve Training with 100% of the data
  - with/without abstract (long text)

Multi-label Classification Problem with Paper Title+Abstract

Data

- paperswithcode: contains **50k** papers from A.I. Field
- ~~arxiv: contains 1.7M papers from many fields – will use A.I. related only~~

Task

- A.I. Field Multi-label Classification
- Active Learning with Labeled Pool (paperswithcode) and Unlabeled Pool (arxiv)

Details

- Uncertainty-based sampling methods
  - Random / Least Confidence / Margin of Confidence / BALD / ~~BatchBALD~~
- Uncertainty-estimation methods
  - Monte Carlo Dropout / ~~Full Ensemble~~ **(Due to HW limits)**
- and compare with Naïve Training with 100% of the data **+ 1%, 5%, 10%, 20%, 50%**
  - ~~with~~/without abstract (long text) **(Again, HW limits)**

Research Questions

Implementing Active Learning from scratch

How much initial data should be given?

Do deep learning models should be **totally "retrained"** every after acquisition?

- For pretrained models, it's okay to start back from the acquisition, but what about others?

Experiments

To discover above questions, I setup the experiments with some models and varying configurations.

Models with their parameters about the similar scale (20~30M)

- **Bidirectional-LSTM** 2-layer, bidirectional.
- **1dCNN-LSTM** 4-layer with varying kernel size.
- **BERT** 4-layer, FF 768-dim, no pretrained models.

Embedding matrix (vocab_size, embed_size) = (83931, 256) ~ 21M

**Detailed methods will be discussed later, if you have any questions.*

Pipeline

1. Make baseline results with varying portion of data
   - ✓ Find the optimal model configuration.
2. Run with varying configurations below.
   - ✓ Acquiring amount : 1%, 5% of total
   - ✓ Acquisition method : Random, LC, Margin, BALD, BatchBALD
   - ✓ Retraining : True or False. When True, acquire when model saturates, otherwise with fixed steps.
   - ✓ Models : BiLSTM, 1dCNN, BERT

Notice

To fit the presentation in time, I omit the followings and will be explained in backup

- ✓ Detailed model setups
- ✓ BERT Results
- ✓ BatchBALD

Setting Model Size & Initial Data Point

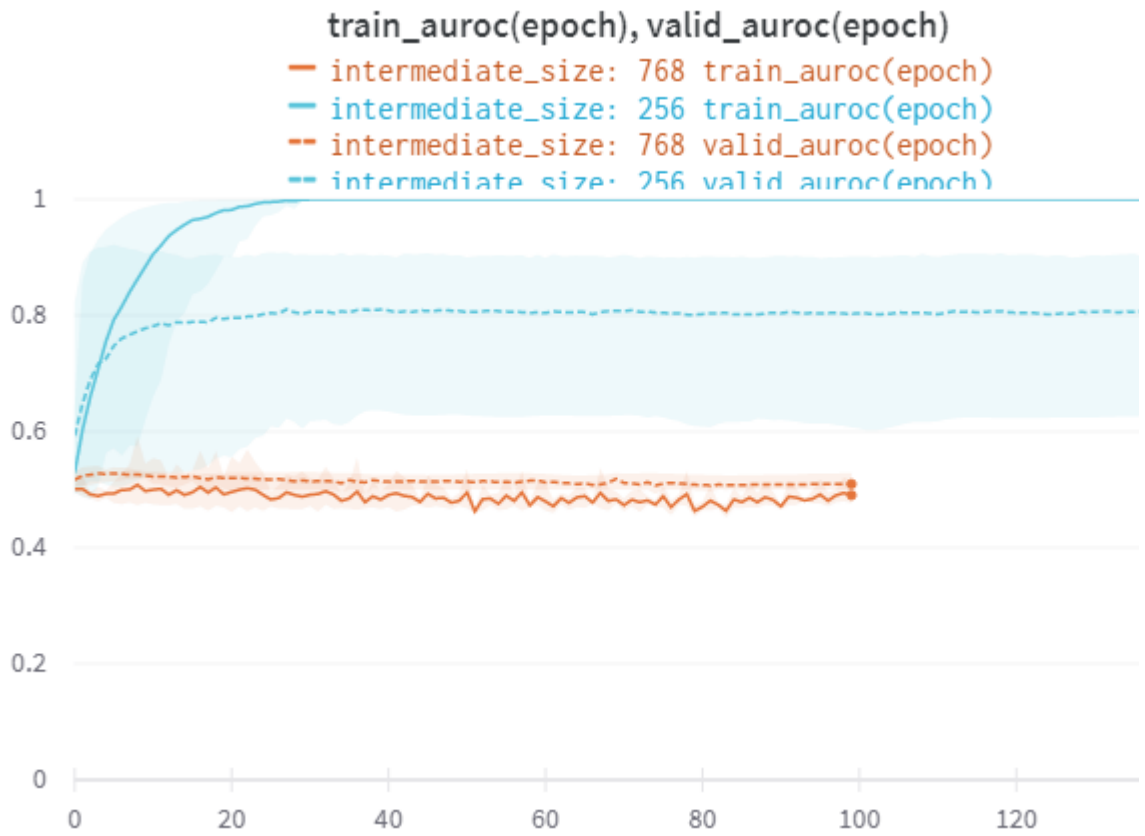
1. Tried with base/large configuration.
2. Experiment with different initial data point

Configuration	size	1dCNN	BiLSTM	BERT
# embed size	Base	256		
	Large	512		
# layers	Base	4	4	4
	Large	4	4	6
# intermediate size (channel)	Base	256	256	256
	Large	1,024	512	768
# params	Base	**25M~**		
	large	**65M~**		

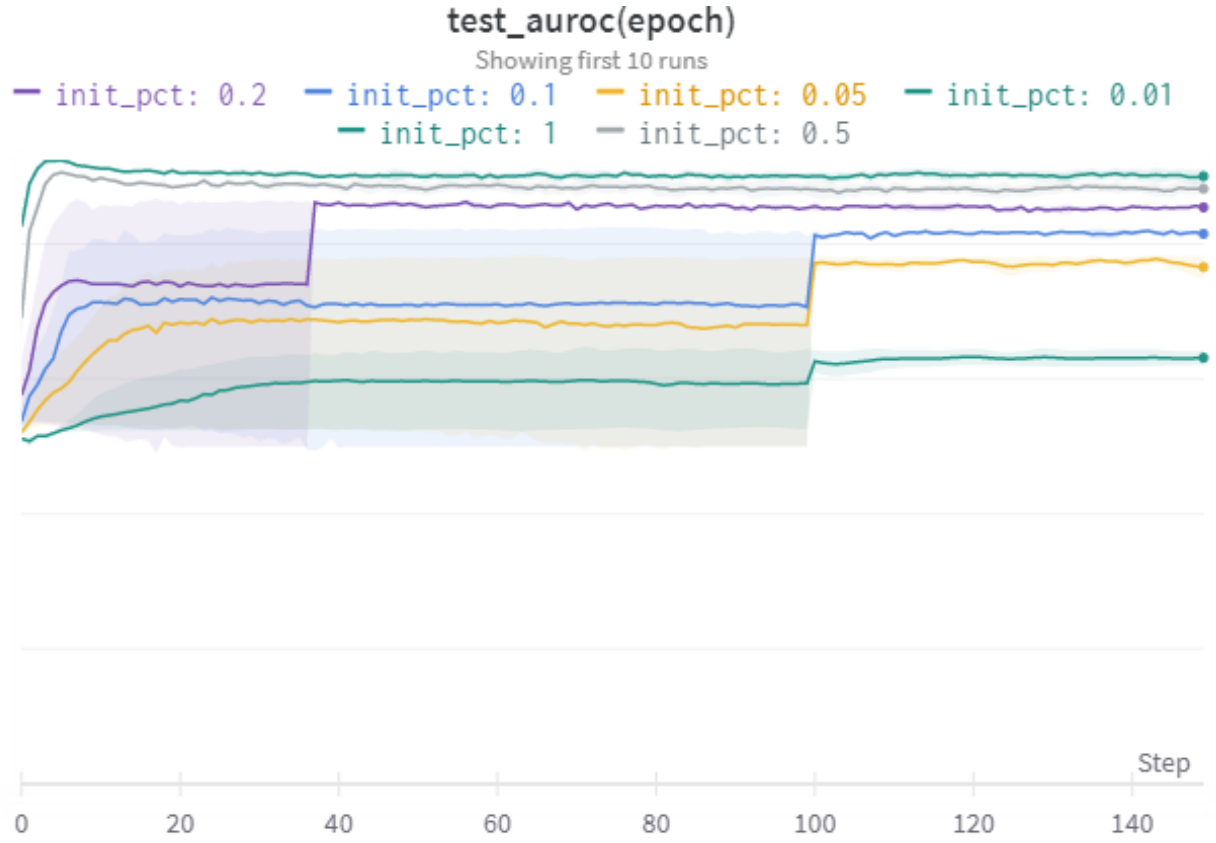
[Table 1] Model Configuration.

Setting Model Size & Initial Data Point

- ✓ Large models can't converge. Used **Base Model**
- ✓ Initial percentage of **20%** was selected.



[Fig 1a] Large vs. Base Model AUROC



[Fig 1b] Initial data point, test AUROC

Naïve Training Baseline Results

BiLSTM 0.8467 | 1dCNN 0.849 | BERT 0.7613

Testing

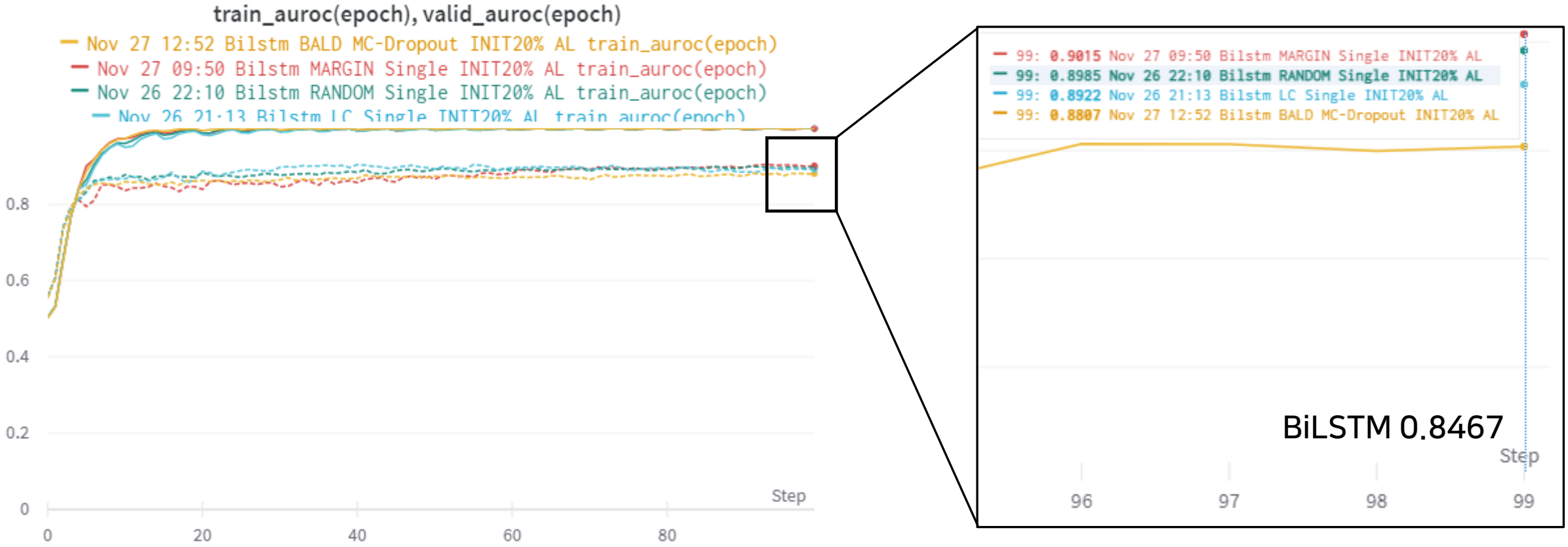
1. Tried with base/large configuration.
2. Experiment with different initial data point

Question	Variation
Let's do Active Learning	Random vs. LC, Margin, BALD
How acquisition period matters?	5, 10 epoch
How much to acquire	1%/5% of total (becomes 500/1,500 per acquisition)
Re-train vs. Keep-train	Compare final test AUROC Re-train and Keep-Train
In overall, did active learning help training?	

[Table 2] Experiment Configuration

Compare Different Uncertainty Selection. (BiLSTM)

- ✓ LC, BALD did not surpass Random, but the difference is not significant.
- ✓ Naïve training setup with 50/100% of the data, 90.32/88.01% AUROC.

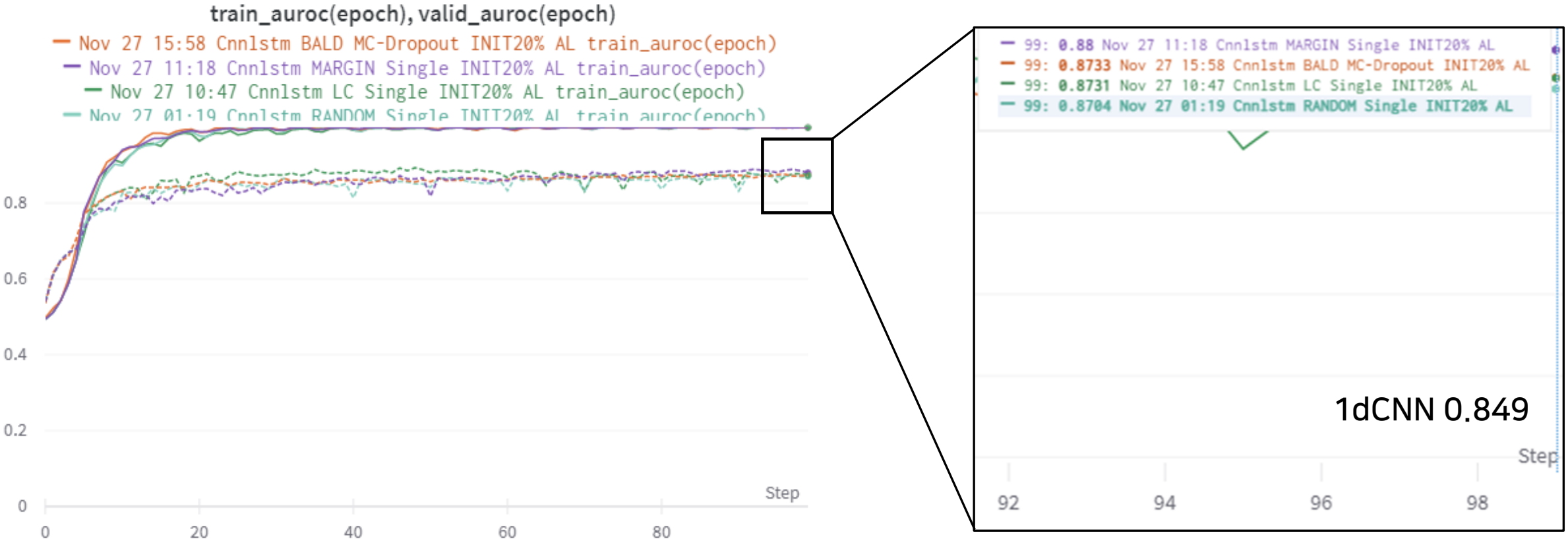


[Fig 2a] Random, LC, Margin, BALD of BiLSTM(base)

Compare Different Uncertainty Selection. (1dCNN)

- ✓ All methods surpass the random,

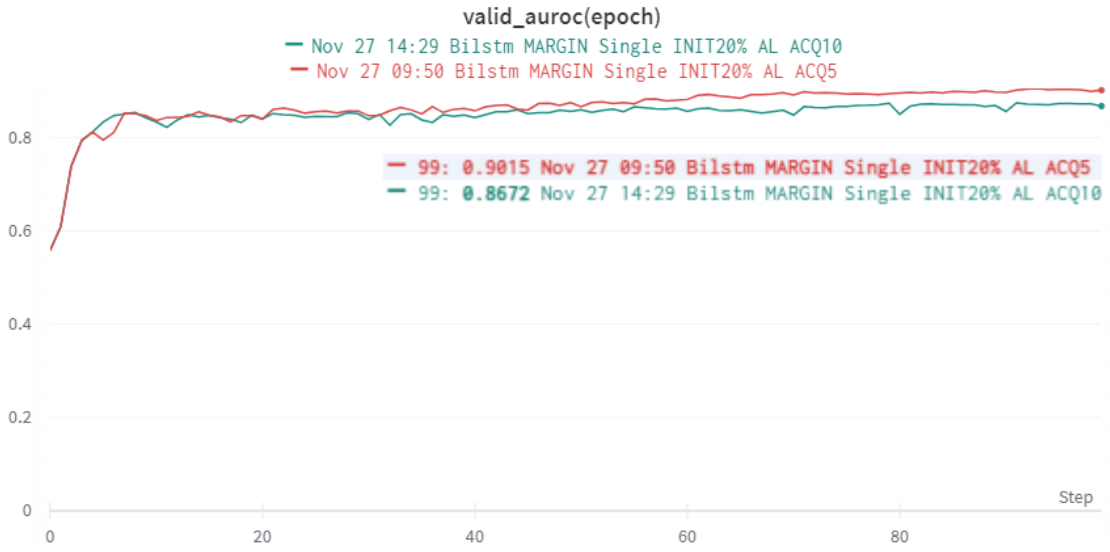
  but only margin looks superior to random in 1%p enhancement.
- ✓ Since with 50% of the total data in naïve training got 87.57% AUROC at last, there was a slight improvement



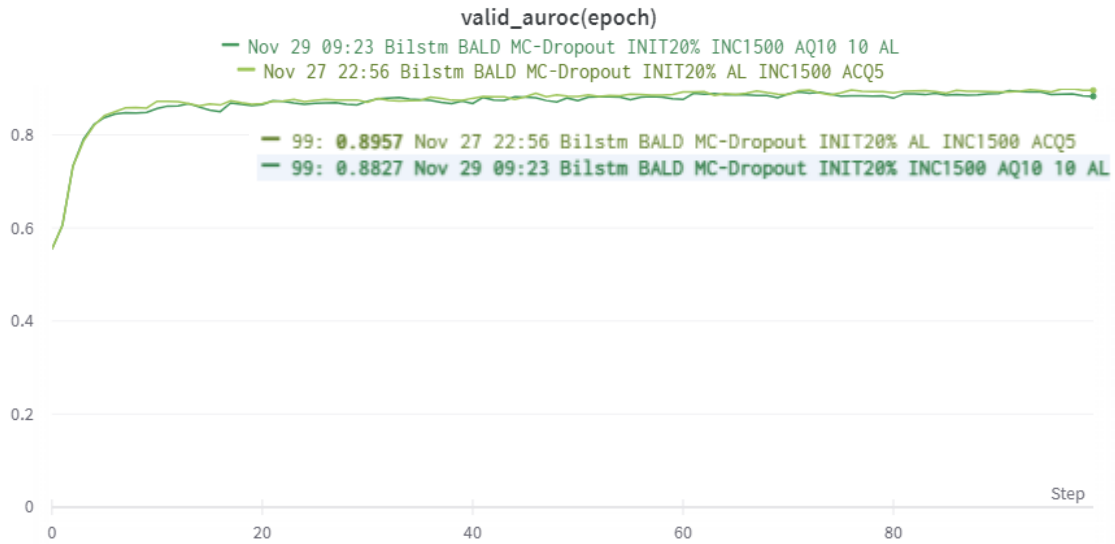
[Fig 2b] Random, LC, Margin, BALD of 1dCNN(base)

Acquisition Period

- ✓ Doesn't have significant difference in BALD
- ✓ Margin showed a slight improvement



[Fig 3a] Acquisition Period 5/10 with BiLSTM Margin

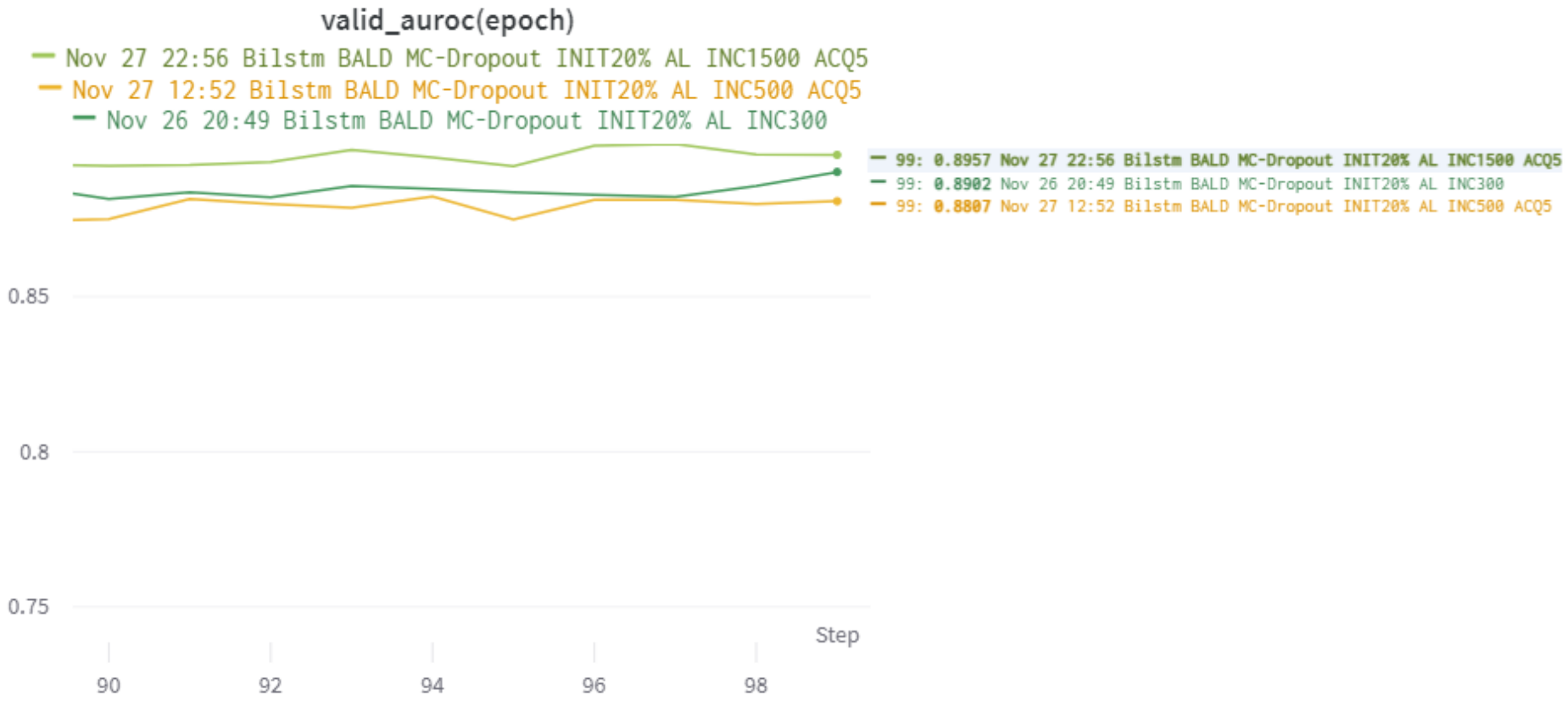


[Fig 3b] Acquisition Period 5/10 with BiLSTM BALD

Compare Increment Amount (BALD, BiLSTM) – 300, 500, 1500

- ✓ As expected, more data leads to better performance.

  (300 vs. 1500 with ~1.5%p improvement)



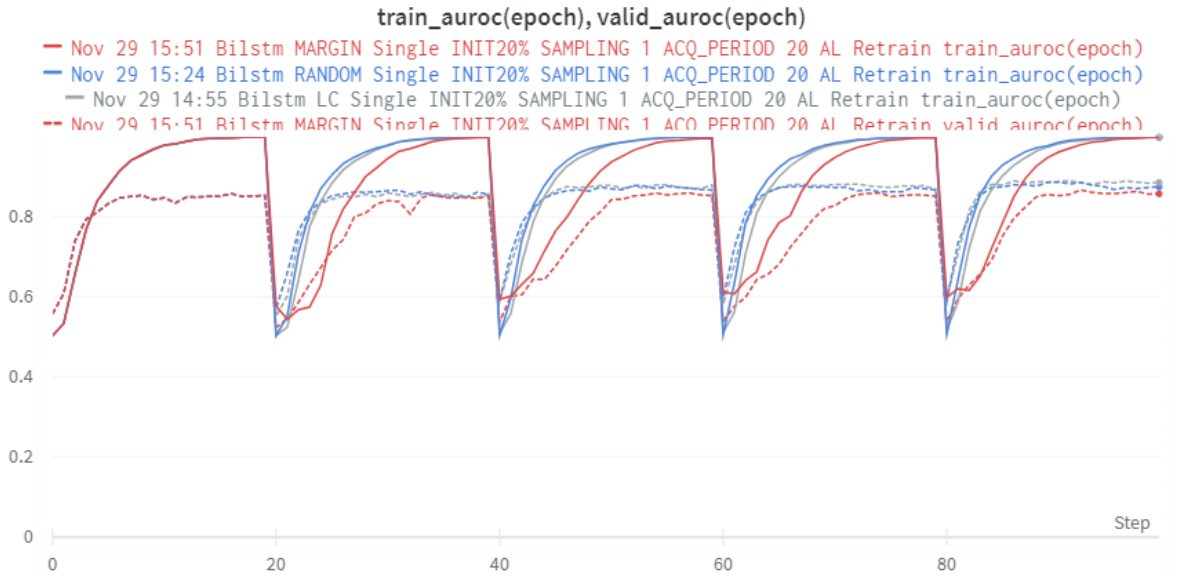
[Fig 4] Increment amount comparison with BiLSTM BALD

Re-train vs. Keep-train

- ✓ Machine learning methods, with the rise of computing power, mostly takes a few seconds to fit to the given training data at least in 50k~100k rows of data.
- ✓ Deep learning methods still needs lots of epochs for model to get trained, and time consuming.
- ✓ In context of active learning, deep learning models get re-trained (re-instantiate and then train) in general even with this exhaustive setup, at least in the project I have seen.
- ✓ In came to my attention that re-training didn't seem to be efficient, so I compared two configurations.

- ✓ For re-training, 1 epoch – 1 acquisition strategy was not plausible for me.
- ✓ Therefore, I tried 20 epoch – 1 acquisition strategy to fit the model.

Re-train vs. Keep-train



[Fig 4a] Retrain with 4 acquisition methods - BiLSTM

[Fig 4b] Retrain with 4 acquisition methods – 1dCNN

Re-train vs. Keep-train

- ✓ LC showed the best performance at last in both models
- ✓ Fluctuations

Model	Method	Valid AUROC step before Acquisition+Retrain				
		20	40	60	80	100
BiLSTM	Random	.8529	.8559 (+.0030)	.8628 (+.0069)	.8667 (+.0039)	.8747 (+.0120)
	LC		**.8567** (+.0038)	**.8797** (+.0023)	**.8674** (+.0123)	**.8853** (+.0179)
	Margin		.8524 (-.0005)	.8542 (+.0018)	.8511 (-.0031)	.8573 (+.0062)
1dCNN	Random	.8221	**.8404** (+.0183)	**.8628** (+.0194)	.8509 (-.0119)	.8608 (+.0099)
	LC		.791 (-.0311)	.8579 (+.0669)	**.8557** (-.0022)	**.8734** (+.0177)
	Margin		.8327 (+.0106)	.8317 (-.00 10)	.8572 (+.0255)	.8513 (-.0059)

[Table 3] Random, LC, Margin / BiLSTM, 1dCNN retrain result organized.

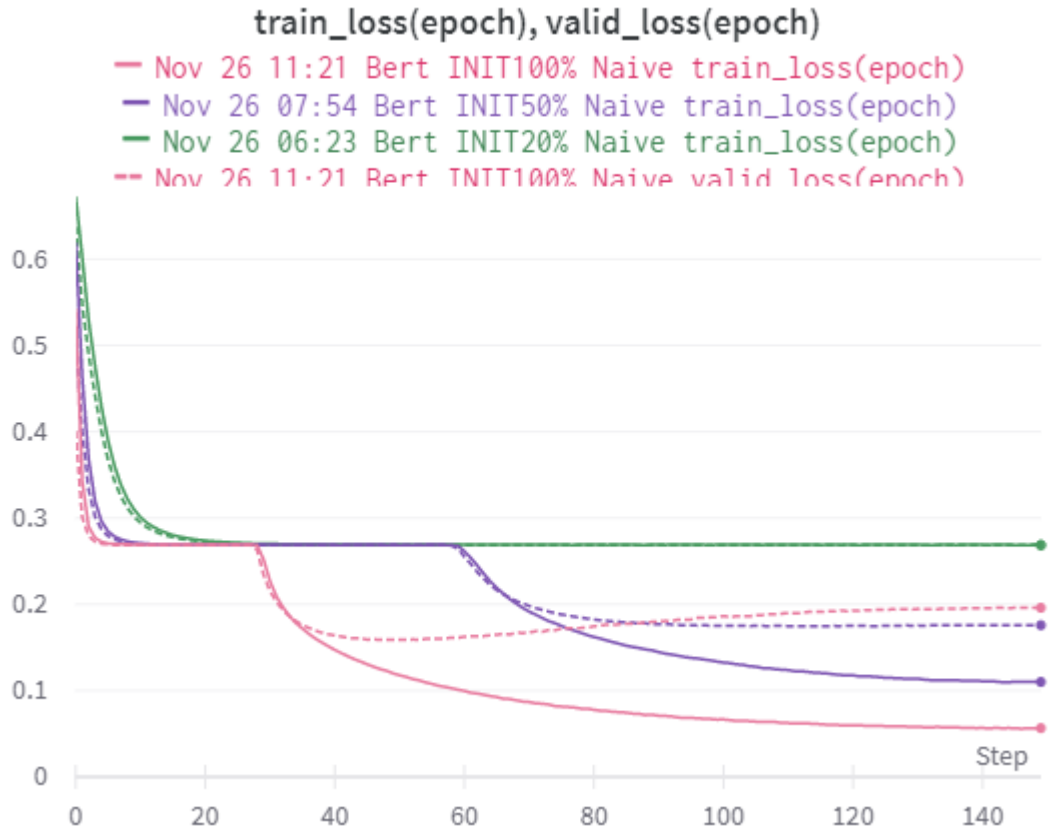
Transformers.

Unlike recurrent neural networks, transformer networks is hard to get trained in this dataset.

Reason is that

- ✓ Not pretrained
- ✓ Shallow network with attention

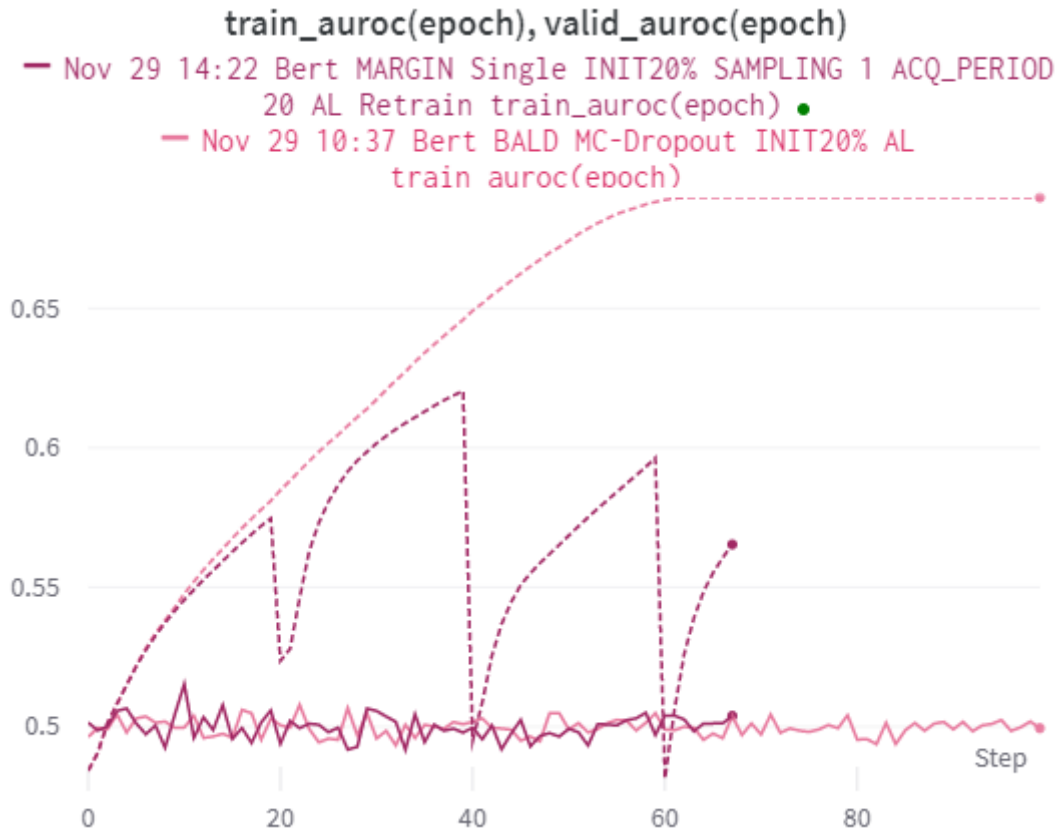
Even the naïve training takes much time to get stabilized.



[Fig 5a] BERT Naïve Training

Transformers.

Tried many different methods… but did not give any insights.

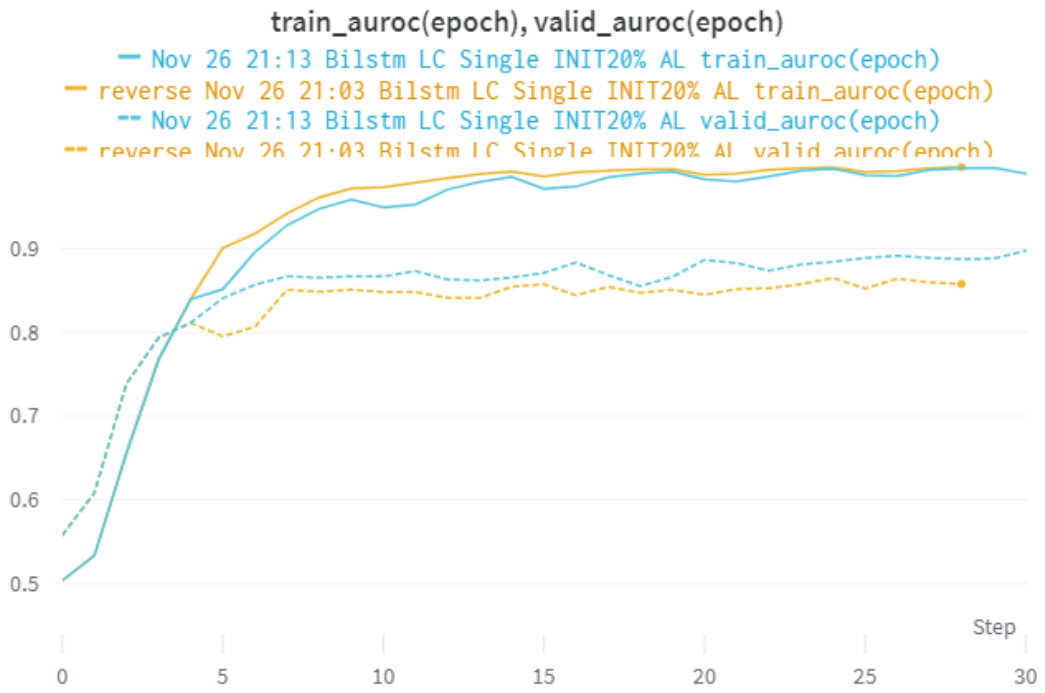


[Fig 5b] Active Learning with BERT

Unintended Ablation Study.

- ✓ We should use high uncertainty samples for acquisition
- ✓ I accidentally sorted these samples in a reverse way,
  so that the samples with lowest uncertainty were chosen
- ✓ Even though I find active learning hard to apply in useful way,
  in this point of view active learning is working

*I know that we need extra runs for validity.



[Fig 6] LC with the most and the least - BiLSTM

Overall, Did Active Learning helped?

- ✓ Some result gives higher performance than random

  but not dramatically high as in related works.

  - ✓ Also some results shows similar/higher performance with the result of more data.
- ✓ Similar performance in the full naïve training setup,

  we can see that active learning not only works in typical toy data **but in real data** as well.

Might be better if …

1. Hardware limits
   - ✓ Train with abstract
   - ✓ Pretrained-transformer models.
   - ✓ Full-ensemble models
2. Extra seeds per trial
3. Plot Uncertainty distribution.

References

About Active Learning

- https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf?sequence=1
- https://dsgissin.github.io/DiscriminativeActiveLearning/2018/07/05/AL-Intro.html
- https://jacobgil.github.io/deeplearning/activelearning

Discriminative Active Learning

- https://arxiv.org/pdf/1907.06347.pdf
  - https://dsgissin.github.io/DiscriminativeActiveLearning/2018/07/05/DAL.html
  - https://github.com/dsgissin/DiscriminativeActiveLearning
- https://kmhana.tistory.com/12?category=838050

BALD

- https://arxiv.org/pdf/1112.5745.pdf
- https://arxiv.org/pdf/1703.02910.pdf
  - https://github.com/Riashat/Deep-Bayesian-Active-Learning

BatchBALD

- https://arxiv.org/abs/1906.08158
  - https://oatml.cs.ox.ac.uk/blog/2019/06/24/batchbald.html
  - https://github.com/BlackHC/BatchBALD

Thank you 🙏
Questions ?

All works+references+detailed experiment resides here.
- ✓ https://github.com/1pha/BayesianActiveLearning
- ✓ https://wandb.ai/1pha/Active-Learning

Daehyun Cho
Cogsys Lab, AI Dept
1phantasmas@korea.ac.kr

Backups & Methods

__unused slides

Multi-label Classification Problem with Paper Title+Abstract

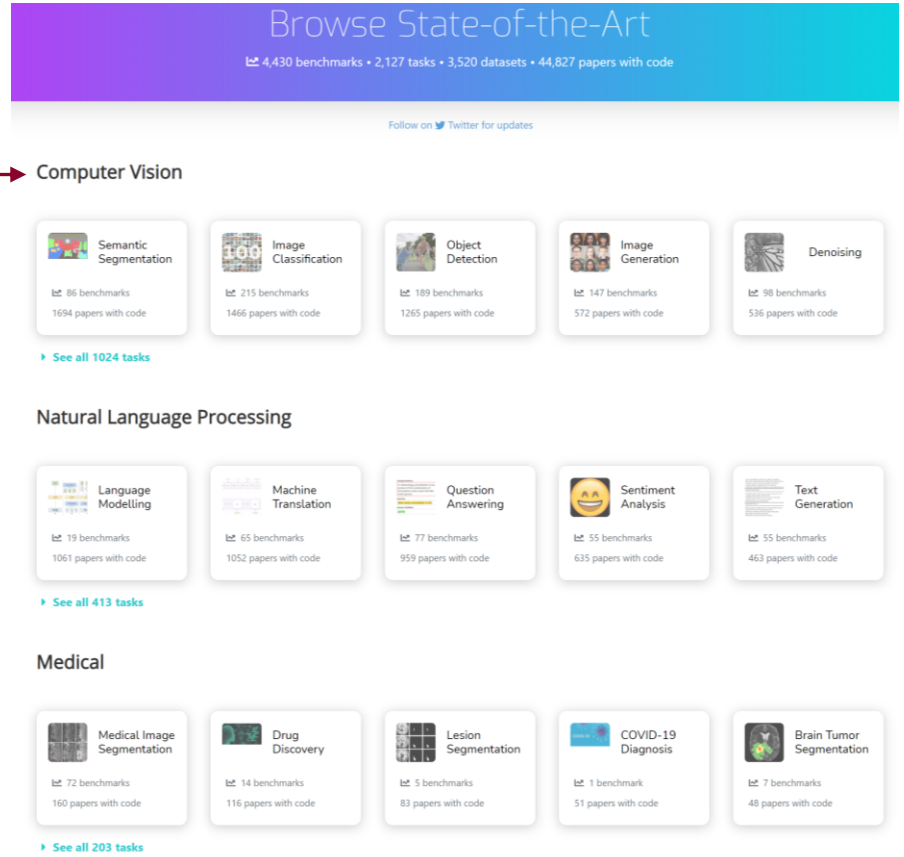
paperswithcode contains 12k papers from A.I. Field

- Contains "specific" A.I. Field
- Some belongs to multiple areas.
- Train/Valid/Test splits with 90/5/5

arxiv contains 1.7M papers from many fields

- Does not contain A.I. Field
- I will use this as unlabeled pool
  - which means … I, oracle, has to do some labe-our

area →



Multi-label Classification Problem with Paper Title+Abstract

Pipeline 1

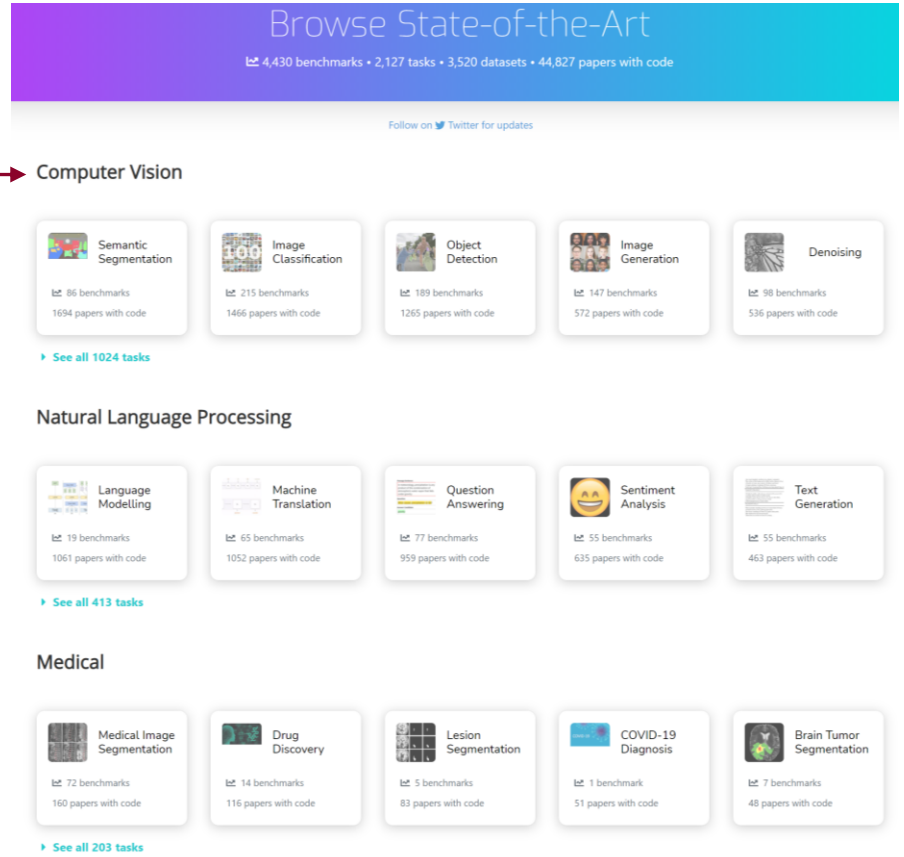
paperswithcode contains 12k papers from A.I. Field

- Contains "specific" A.I. Field
- Some belongs to multiple areas.
- Train/Valid/Test splits with 90/5/5

arxiv contains 1.7M papers from many fields

- Does not contain A.I. Field
- I will use this as unlabeled pool
  - which means … I, oracle, has to do some labe-our

area →



Alumni

__slides from Interim Presentation

Multi-label Classification Problem with Paper Title+Abstract

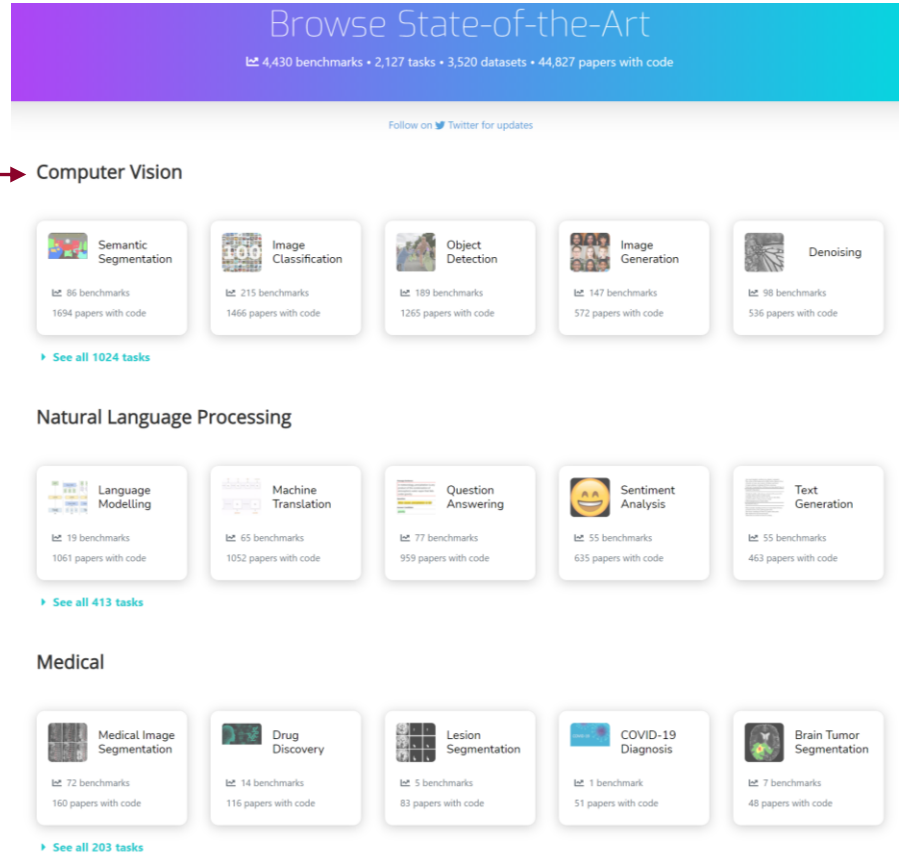
paperswithcode contains 12k papers from A.I. Field

- Contains "specific" A.I. Field
- Some belongs to multiple areas.
- Train/Valid/Test splits with 90/5/5

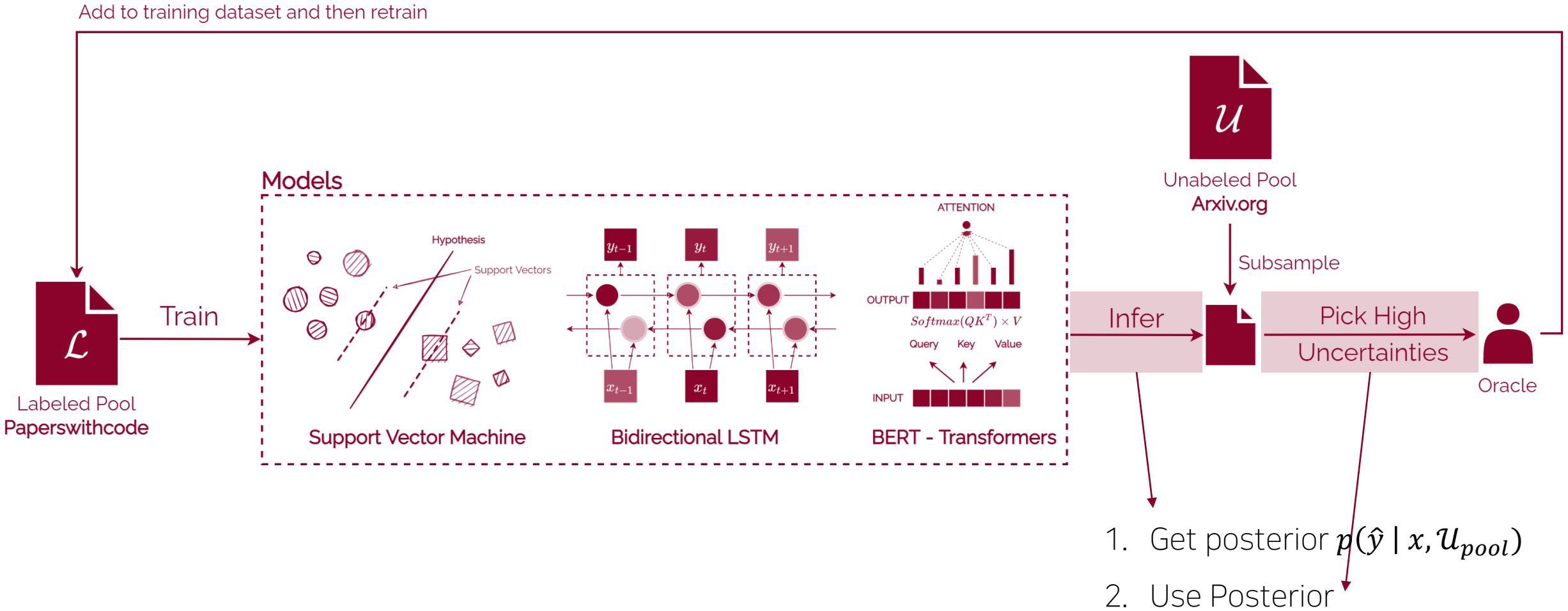
arxiv contains 1.7M papers from many fields

- Does not contain A.I. Field
- I will use this as unlabeled pool
  - which means … I, oracle, has to do some labe-our

area →



Whole Pipeline in Blueprint



How to get Uncertainties / Posteriors ? (Inference)

Not just the prediction, but "posterior probability" of the predictions. $p(\hat{y} \mid x, \mathcal{U}_{pool})$

Dropout as an approximation [1]

- Applied dropout can is equivalent to an approximation to the probabilistic deep Gaussian Processes and minimizes KL-divergence with the posterior [1]
- To simply put it, $p(\hat{y} \mid x, \mathcal{U}_{pool}) = \frac{1}{T}\sum p(\hat{y} \mid x, w_t)$, sum of many different dropout models [3]

Ensemble models [2, 3]

- Trained multiple models and compared with the method above
- Ensembling 5 models surpassed T=25 from MC in performance (MNIST, CIFAR-10)
  - They view this problem as same weights/initialization/optimization in MC Dropout [3-4.3]

[1] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.
[2] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *arXiv preprint arXiv:1612.01474* (2016).
[3] Beluch, William H., et al. "The power of ensembles for active learning in image classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

How to alleviate Uncertainties ?: Classic Active Learning [4]

Uncertainty Sampling

- Random Sampling
- Least Confidence
- Margin Sampling

Query-By-Committee (QBC)

- Vote Entropy
- KL-Divergence

Expected Model Change (EGL)

- Vote Entropy
- KL-Divergence

Variance Reduction and Fisher Information Ratio (FIR)

[4] Settles, Burr. "Active learning literature survey." (2009).

How to alleviate Uncertainties ? - for Deep Learning

Deep Learning has difficulties in - [5]

Requiring Large amounts of data

No representation about model uncertainty

Discriminative Active Learning (DAL) [6]

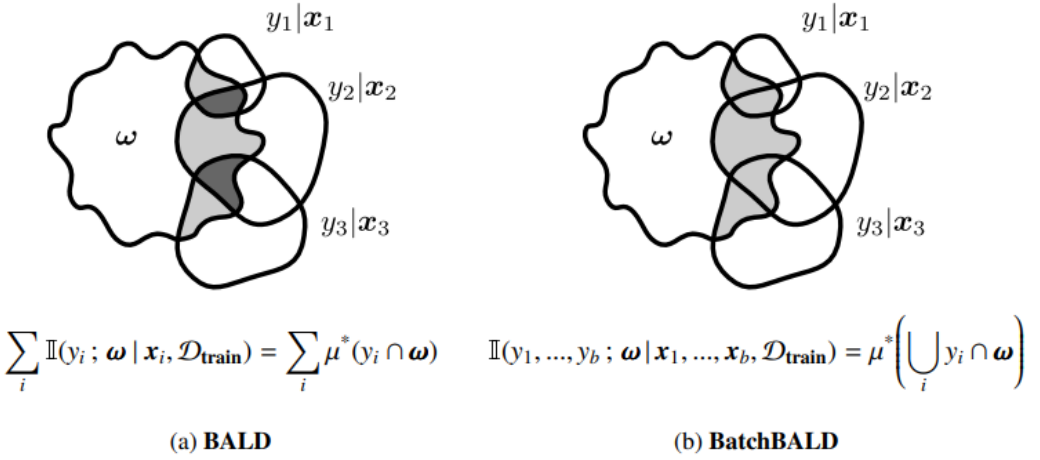
- Select a sample that is far from learned representation

Bayesian Active Learning by Disagreement (BALD) [7]

- Find examples whose output is marginally uncertain, with many disagreements between sampled models

BatchBALD [8]

- Tackled problem of selecting “near” duplicates from BALD



[5] Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." *International Conference on Machine Learning*. PMLR, 2017.
[6] Gissin, Daniel, and Shai Shalev-Shwartz. "Discriminative active learning." *arXiv preprint arXiv:1907.06347* (2019).
[7] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." *arXiv preprint arXiv:1112.5745* (2011).
[8] Kirsch, Andreas, Joost Van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning." *Advances in neural information processing systems* 32 (2019): 7026-7037.

Applications to Natural Language Processing tasks ?

1dCNN / LSTM / CRF + AL

- LC / MNLP / BALD Sampling on NER – SOTA with much less data [9]
- LC / Dropout+BALD / Backprop-by-Bayes on few tasks [10]
  - SC: no significance / NER, SRL: with 50% of the dataset, outperforms w/o AL

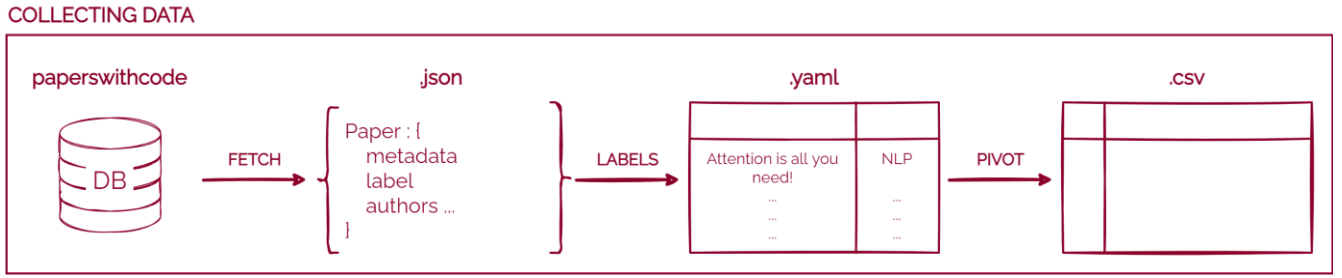
BERT + AL

- BERT Finetune on Unlabeled Pool gives performs better than standard BERT Finetuning [11]
- In real-world challenging scenario, AL can improve model performance [12]
- BERT Classification task with AL performs well [13]
- No single strategy outperforms the other [12, 13]

[9] Shen, Yanyao, et al. "Deep active learning for named entity recognition." *arXiv preprint arXiv:1707.05928* (2017).
[10] Siddhant, Aditya, and Zachary C. Lipton. "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study." *arXiv preprint arXiv:1808.05697* (2018).
[11] Margatina, Katerina, Loic Barrault, and Nikolaos Aletras. "Bayesian Active Learning with Pretrained Language Models." *arXiv preprint arXiv:2104.08320* (2021).
[12] Dor, Liat Ein, et al. "Active learning for BERT: An empirical study." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
[13] Prabhu, Sumanth, Moosa Mohamed, and Hemant Misra. "Multi-class Text Classification using BERT-based Active Learning." *arXiv preprint arXiv:2104.14289* (2021).

Paperswithtopic Data Collection Pipeline

Experiments



Use paperswithcode Database

- Database page made with Django library with their domain open
- Everyone can fetch data here!
  - There is an open API made by them, but does not work
  - I partially used their open-source code to scrape the data

Fetching and Organizing Data

- Right figure is the raw meta-data fetched from the DB
- Here I only used 'title' and 'area'
  - 'area' is not seen on the figure since papers were scraped by area

```
1 paper_meta_dict['"Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Po
executed in 6ms, finished 08:26:14 2021-04-29
{'id': 'brilliant-ai-doctor-in-rural-china-tensions',
 'arxiv_id': '2101.01524',
 'nips_id': None,
 'url_abs': 'https://arxiv.org/abs/2101.01524v2',
 'url_pdf': 'https://arxiv.org/pdf/2101.01524v2.pdf',
 'title': '"Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Powered CDSS
Deployment',
 'abstract': 'Artificial intelligence (AI) technology has been increasingly used in the imp
lementation of advanced Clinical Decision Support Systems (CDSS). Research demonstrated the
potential usefulness of AI-powered CDSS (AI-CDSS) in clinical decision making scenarios. Ho
wever, post-adoption user perception and experience remain understudied, especially in deve
loping countries. Through observations and interviews with 22 clinicians from 6 rural clini
cs in China, this paper reports the various tensions between the design of an AI-CDSS syste
m ("Brilliant Doctor") and the rural clinical context, such as the misalignment with local
context and workflow, the technical limitations and usability barriers, as well as issues r
elated to transparency and trustworthiness of AI-CDSS. Despite these tensions, all particip
ants expressed positive attitudes toward the future of AI-CDSS, especially acting as "a doc
tor\'s AI assistant" to realize a Human-AI Collaboration future in clinical settings. Final
ly we draw on our findings to discuss implications for designing AI-CDSS interventions for
rural clinical contexts in developing countries.',
 'authors': ['Dakuo Wang',
  'Liuping Wang',
  'Zhan Zhang',
  'Ding Wang',
  'Haiyi Zhu',
  'Yvonne Gao',
  'Xiangmin Fan',
  'Feng Tian'],
 'published': '2021-01-04',
 'conference': None,
 'conference_url_abs': None,
 'conference_url_pdf': None,
 'proceeding': None}
```

Fig. 1 Raw Meta Data for each paper

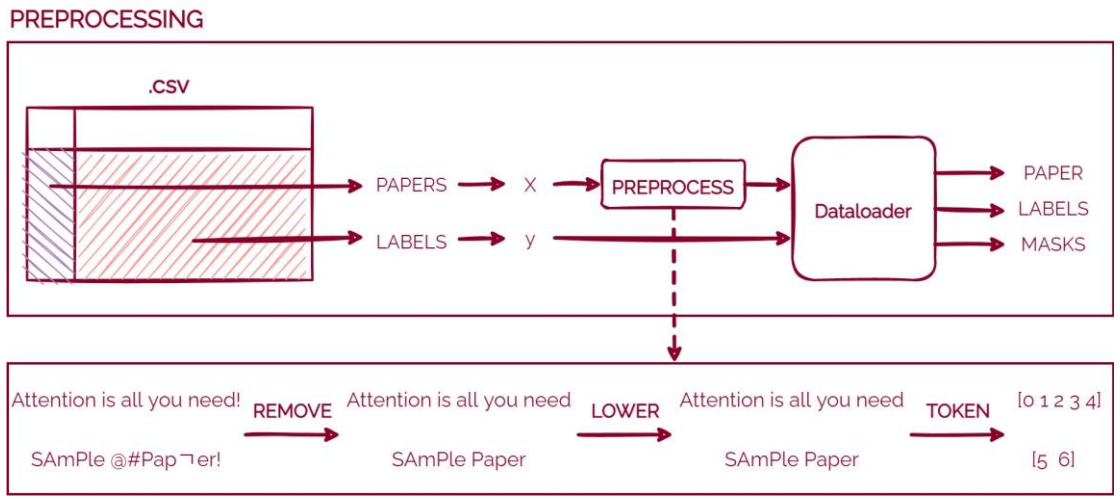
Preprocessing

Experiments

Make correct X, y

- Preprocessing is the most important part in NLP

1. I only leaved Alphabets and Numbers
   - Chinese, Special characters were all removed
2. Lowered the alphabets
3. Tokenized by word
4. Embedding
   - This can be done with many ways, such as –
   - sentencepiece, word2vec, FastText
     - Possible to remove tenses (plural, past tense e.t.c)
   - Word embeddings
     - But we can also expect deep learning to do that as well.



Active Learning (with my humble interpretation)

There must be many ways to do it, but in my case -

1. Train the model (or models) with Training data
2. Infer Unlabeled Pool (may not be whole, but some) to get the posterior

$$p(y = class \mid x, D_{pool})$$

3. With calculated posterior, sample data that has high uncertainty through followings
    1. Uncertainty Sampling
        - Least Confidence, Margin sampling, Entropy Sampling
    2. Query by Committee (through multiple models)
        - Vote Entropy, KL-Divergence, ⋯
        - BALD (but probabilistic), BatchBALD, ⋯
    +. Expected Model Change, Density-based method (Core-set, REPR) , e.t.c.
4. Add these samples to training data and retrain

Why this project?

Why Multi-label Classification?

- It was interesting that AL can do the work with small portion of data
  in many classification tasks by finding decision boundaries with tactic.
- Wondering if **multi-label classification would work** too.

Why NLP?

- There were some works about NLP + AL in the field, but not deeply
  and especially Attention models were hard to find
- Also, I have some side project on this and became curious about how AL would fit in.

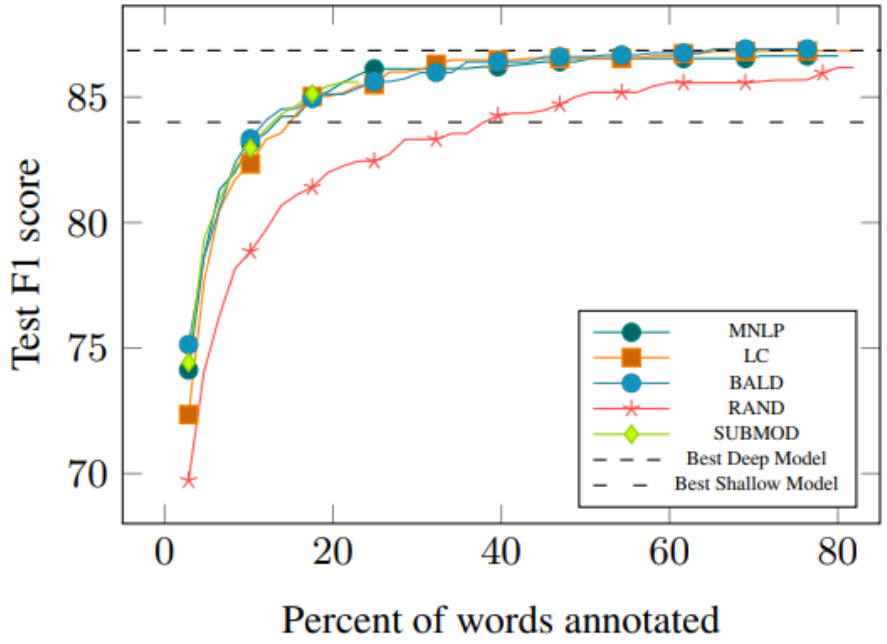
Why SVM / BiLSTM / BERT ?

- ML models screwed up in my previous projects and wanted to see if AL works for SVM.
- BiLSTM was most widely used model in NLP+AL Related works
- BERT easy to implement and definitely should be tell apart with Recurrent Models
  - Other transformer models in candidate as well

Deep active learning for named entity recognition [1]

One of early trials using Active learning in Deep learning models

- Used MNLP/LC [2] sampling methods and BALD [3] and Random as baseline
- This work came out before the boom of Transformers,
  1dCNN and BiLSTM were used
- Would be interesting to find more insight
- tested on CoNLL-2003, OntoNotes-5.0 (2013)
- Achieved SOTA trained with standard methods
  with much less data

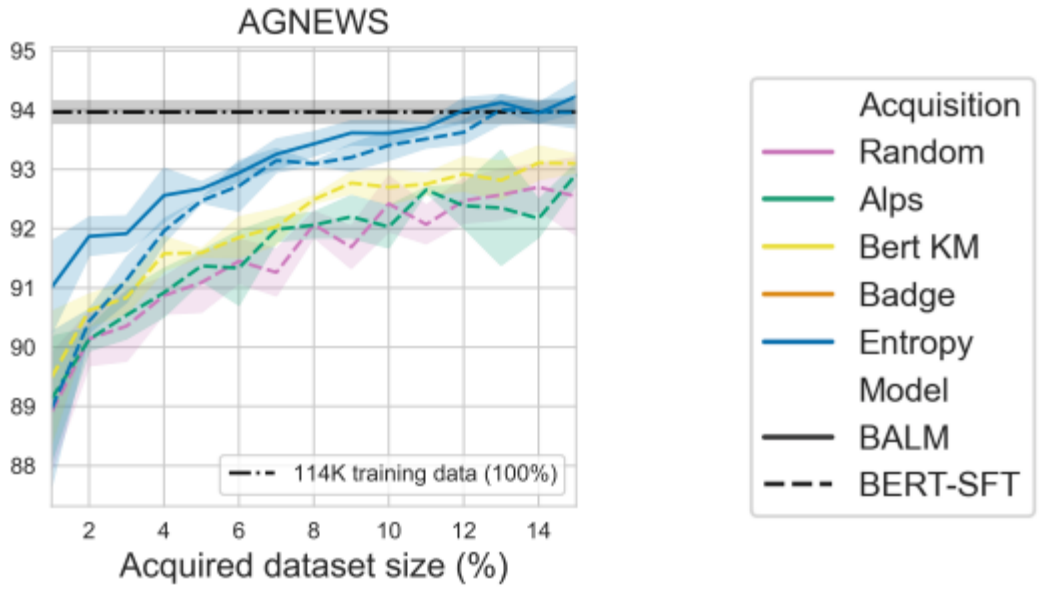


[1] Shen, Yanyao, et al. "Deep active learning for named entity recognition." *arXiv preprint arXiv:1707.05928* (2017).
[2] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." *arXiv preprint arXiv:1112.5745* (2011).
[3] Settles, Burr. "Active learning literature survey." (2009).

Bayesian Active Learning with Pretrained Language Models. [4]

Transformer architecture experiments with multiple AL strategy

- Pretrained Models are Finetuned again in a specific task
- In this finetuning stage, this work used active learning in order to achieve high performance with much less data
- Compared with standard finetuning methods with multiple datasets
- Within 15% of all datasets, active learning strategy surpassed the standard training options.
  - No acquisition strategy universally performs better



[4] Margatina, Katerina, Loic Barrault, and Nikolaos Aletras. "Bayesian Active Learning with Pretrained Language Models." *arXiv preprint arXiv:2104.08320* (2021).

BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning

BALD [5]

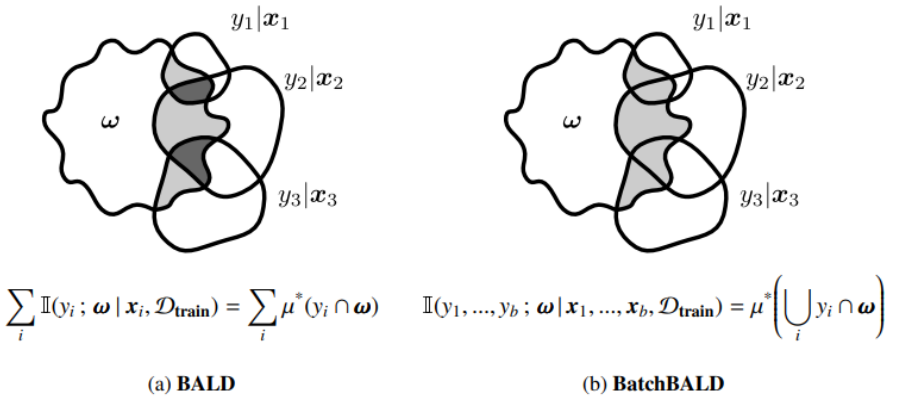
- Find examples whose output is marginally uncertain, with many disagreements between sampled models

$$I(y;\omega|x,D_{train}) = H(y|x,D_{train}) - E_{p(\omega|D_{train})}[H(y|x,\omega,D_{train})]$$

  - Tries to find images with high uncertainty and disagreements on different models
  - Through MC Dropout, we can get the approximation of this. [6]
    - Full-ensemble is used as well [7]

BatchBALD [8]

- Tackled problem of selecting "near" duplicates from BALD
- Reduce redundancy of similar samples being selected



[5] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." *arXiv preprint arXiv:1112.5745* (2011).
[6] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.
[7] Kirsch, Andreas, Joost Van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning." *Advances in neural information processing systems* 32 (2019): 7026-7037.
[8] Beluch, William H., et al. "The power of ensembles for active learning in image classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Through this project …

Similar/Same approach to novel dataset

- More insight about NLP + AL
- Trials with various ML/Transformer models
- Domain-specific dataset

Serving (if possible)

Thinking of data annotation tool

- Annotating “high uncertainty” data first would help in large amounts of data
- Not going to deploy seriously, but just a mockup