Paper Topic Classification with Active Learning

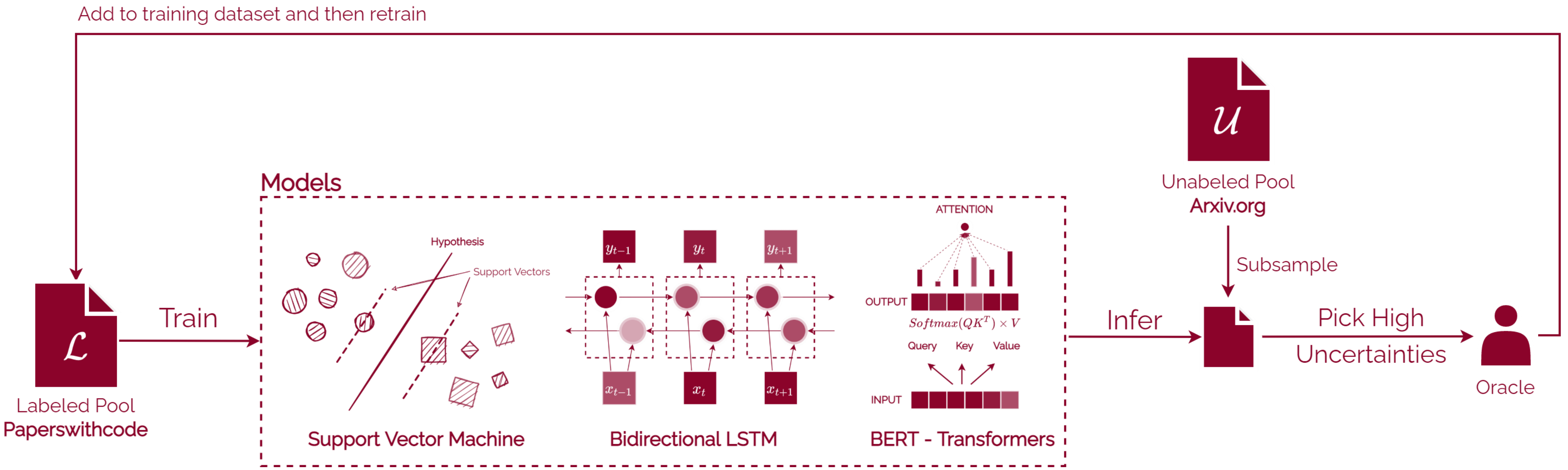
Daehyun Cho
Cognitive System Lab
A.I. Dept, Korea Univ.

Index

Cogsys Lab

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

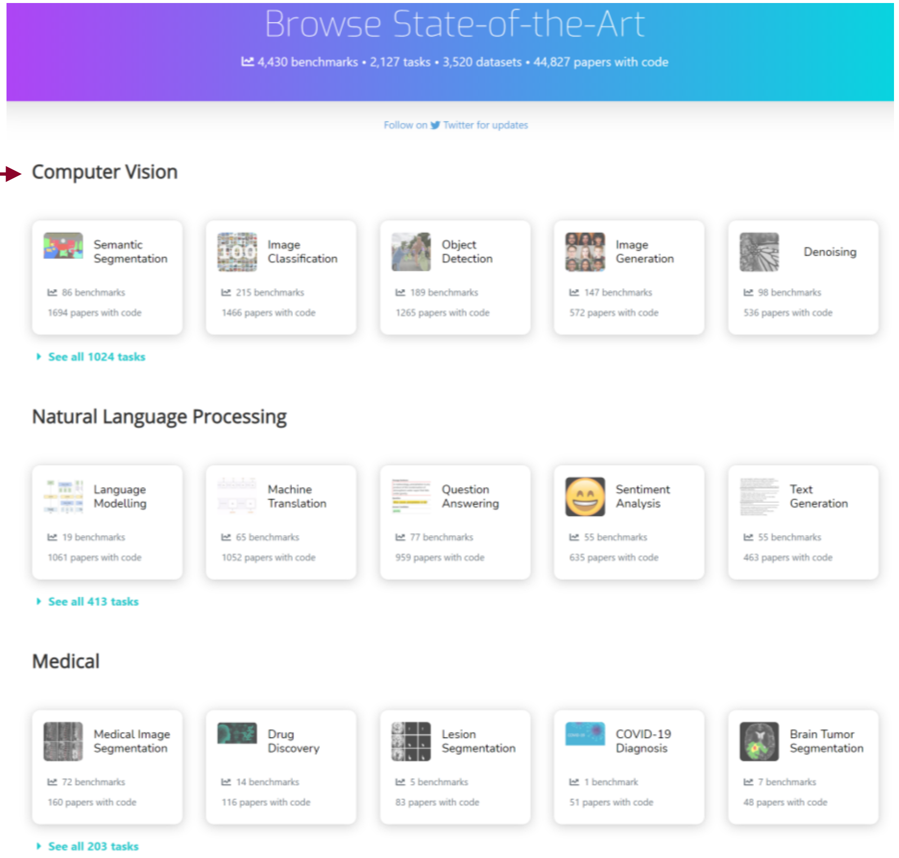
paperswithcode contains 12k papers from A.I. Field

- 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 →



Aims to …

Achieve better results models from previous project on test set

- ML Models didn't work properly before …
- Transformer models were high enough, peak around 94% but wanted to see if how far this can reach with less data

Types	Models	AUROC
Machine Learning	Extra Tree	52.4 ± 0.02
	Complement Naive Bayes	52.5 ± 0.12
	Naive Bayes	52.9 ± 0.05
	K-Nearest Neighbour	53.2 ± 0.03
	Random Forest	56.4 ± 0.05
	AdaBoost	59.8 ± 0.06
	LightGBM	63.9 ± 0.08
	XGBoost	66.8 ± 0.09

Table 1: AUROC comparison between machine learning models, using tokenized input. Error indicates Standard Error of the Mean (SEM) across 10 folds.

Model	#L	#A	#H		
			128	256	512
ELECTRA	1	8	93.0	93.1	92.7
		16	93.1	93.1	**93.6**
		32	93.6	93.4	93.6
	2	8	91.5	92.9	88.4
		16	92.5	93.7	90.8
		32	92.6	**93.8**	92.6
	3	8	*53.7*	93.2	81.3
		16	*53.1*	**93.7**	87.5
		32	91.2	93.1	92.2
	4	8	*52.6*	93.5	74.4
		16	*50.3*	93.3	84.4
		32	*53.7*	**93.7**	89.9

Table 3: Grid search result over number of layers, heads and hidden dimension for the ELECTRA Transformer Model pooled with first sequence label: #L=the number of layers; #A=the number of attention heads; #H=hidden size.

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.

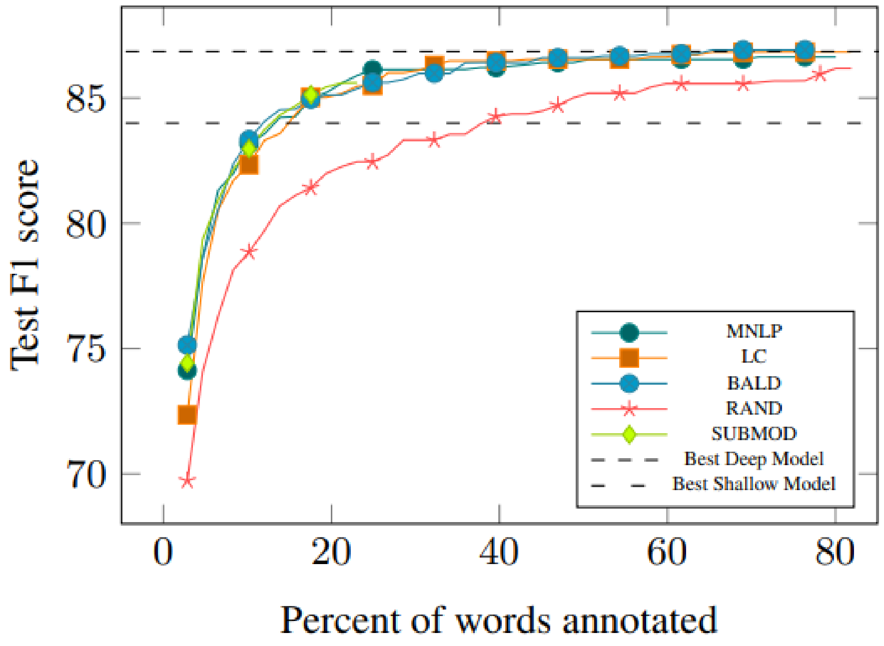
Why SVM / BiLSTM / BERT ?

- 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

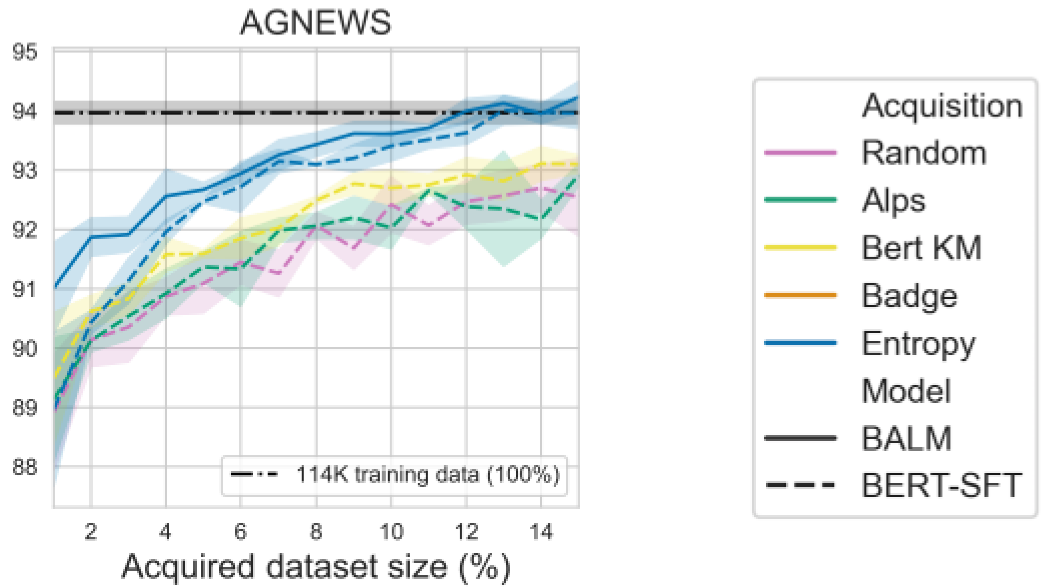


[1] Shen, Yanyao, et al. "Deep active learning for named entity recognition." *arXiv preprint arXiv:1707.05928* (2017).
[2] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." *arXiv preprint arXiv:1112.5745* (2011).
[3] Settles, Burr. "Active learning literature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

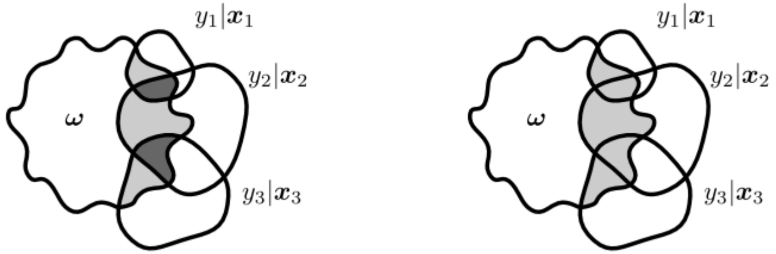
- Find examples whose output is marginally uncertain, with many disagreements between sampled models

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

  - 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



$\sum_i \mathbb{I}(y_i\,;\boldsymbol{\omega}\mid\boldsymbol{x}_i,\mathcal{D}_{\text{train}}) = \sum_i \mu^*(y_i\cap\boldsymbol{\omega})$

(a) BALD

$\mathbb{I}(y_1,\ldots,y_b\,;\boldsymbol{\omega}\mid\boldsymbol{x}_1,\ldots,\boldsymbol{x}_b,\mathcal{D}_{\text{train}}) = \mu^*\left(\bigcup_i y_i\cap\boldsymbol{\omega}\right)$

(b) BatchBALD

[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 …

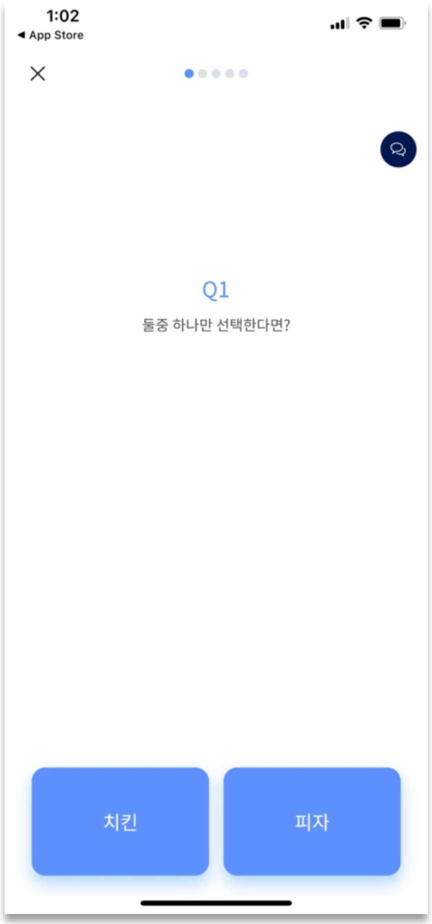
Similar/Same approach to novel dataset

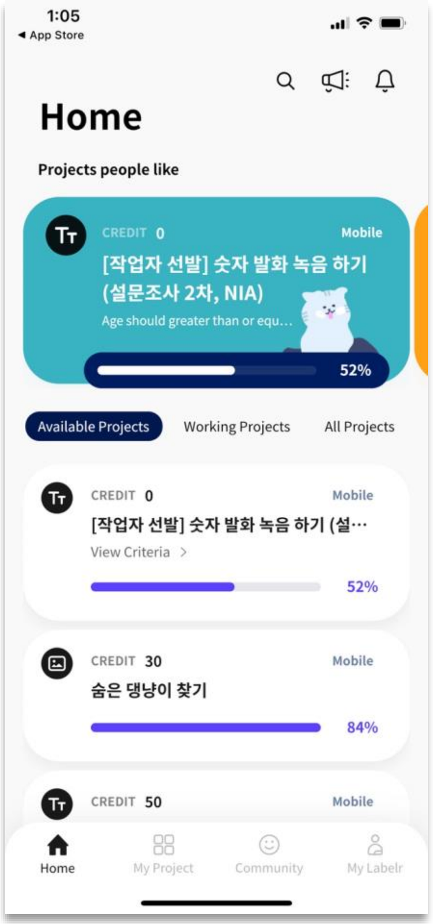
- More insight about NLP + AL
- Trials with various ML/Transformer models
- Domain-specific dataset

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





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

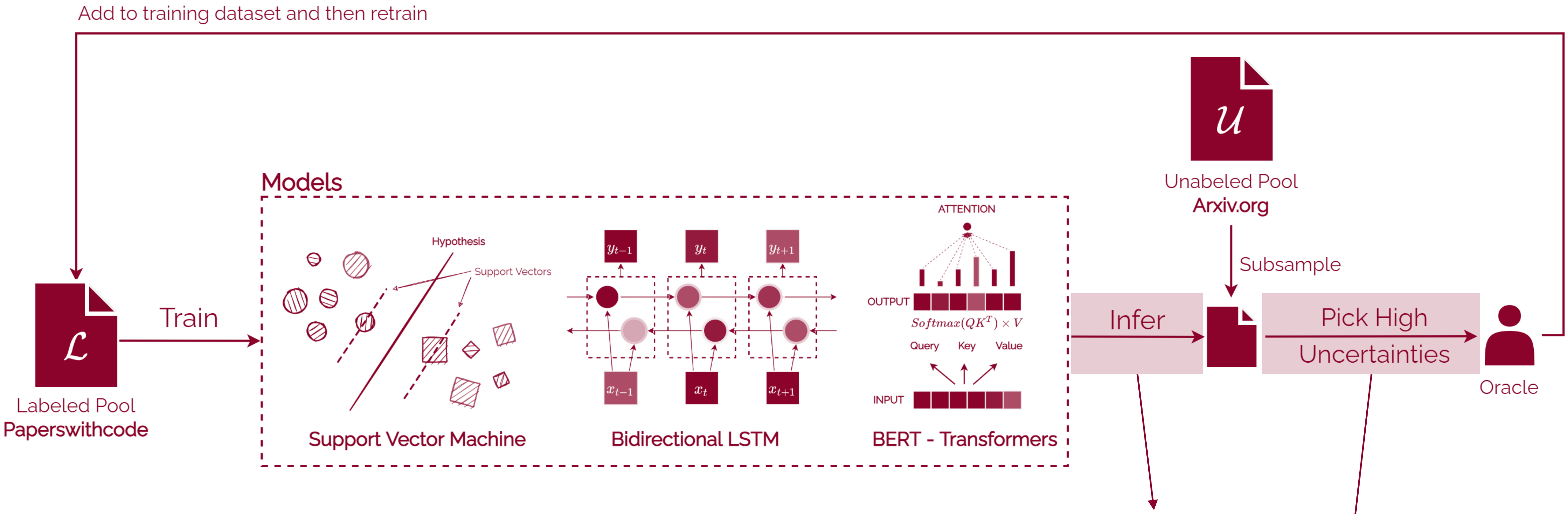
Thank you 🙏
Questions ?

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

Backups

__unused slides

Whole Pipeline in Blueprint



1. Get posterior $p(\hat{y} \mid x, \mathcal{U}_{pool})$
2. Use Posterior

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]

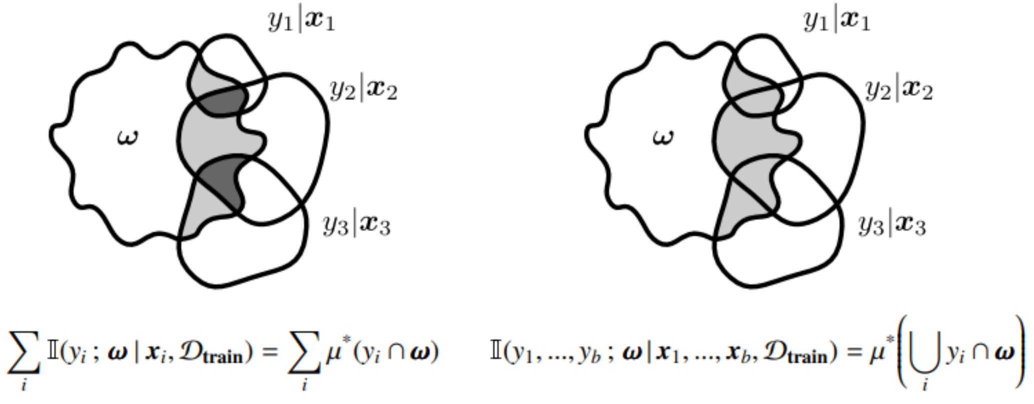
- Select a sample that is far from learned representation

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



(a) **BALD** (b) **BatchBALD**

[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

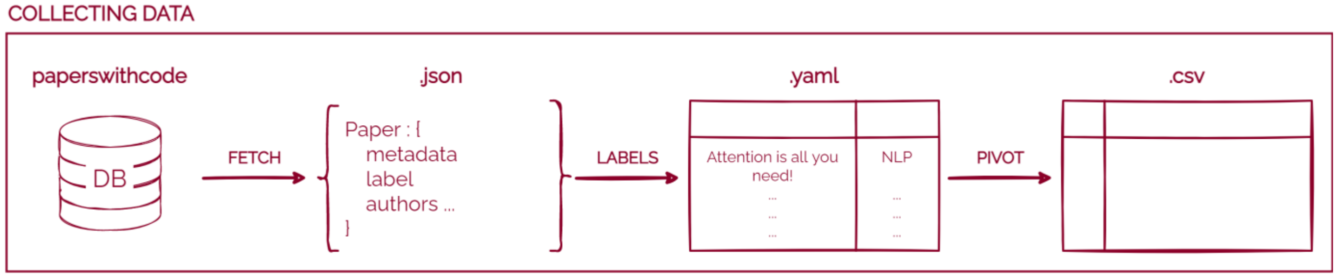
BERT + AL

- 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

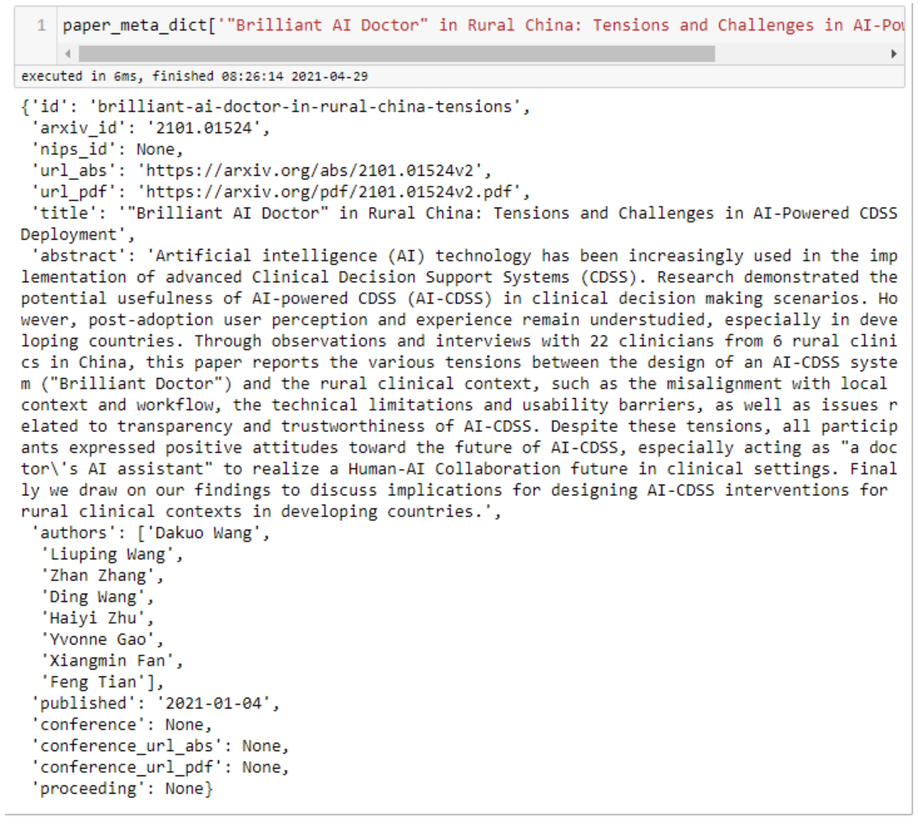


Use paperswithcode Database

- 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




```
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 implementation of advanced Clinical Decision Support Systems (CDSS). Research demonstrated the potential usefulness of AI-powered CDSS (AI-CDSS) in clinical decision making scenarios. However, post-adoption user perception and experience remain understudied, especially in developing countries. Through observations and interviews with 22 clinicians from 6 rural clinics in China, this paper reports the various tensions between the design of an AI-CDSS system ("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 related to transparency and trustworthiness of AI-CDSS. Despite these tensions, all participants expressed positive attitudes toward the future of AI-CDSS, especially acting as "a doctor\'s AI assistant" to realize a Human-AI Collaboration future in clinical settings. Finally 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

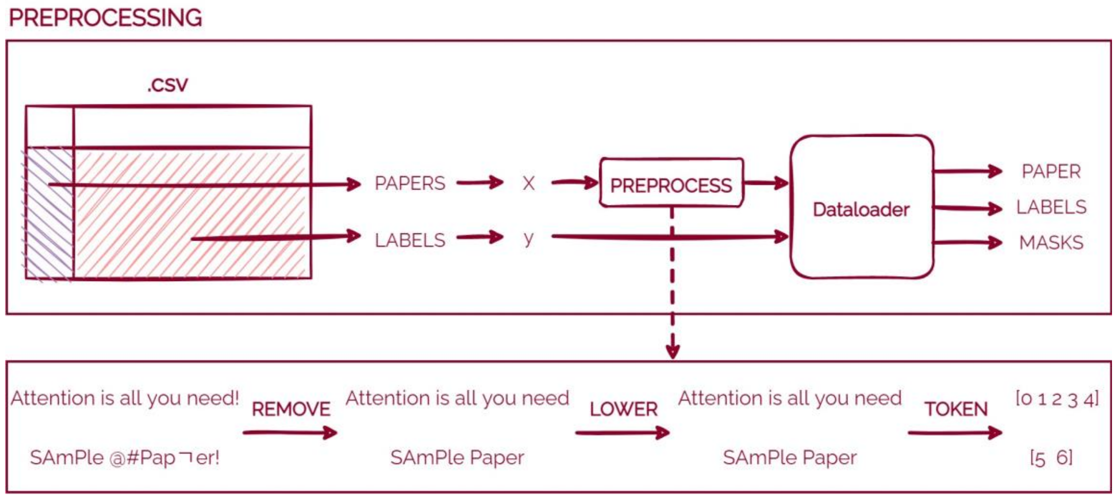
Preprocessing

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