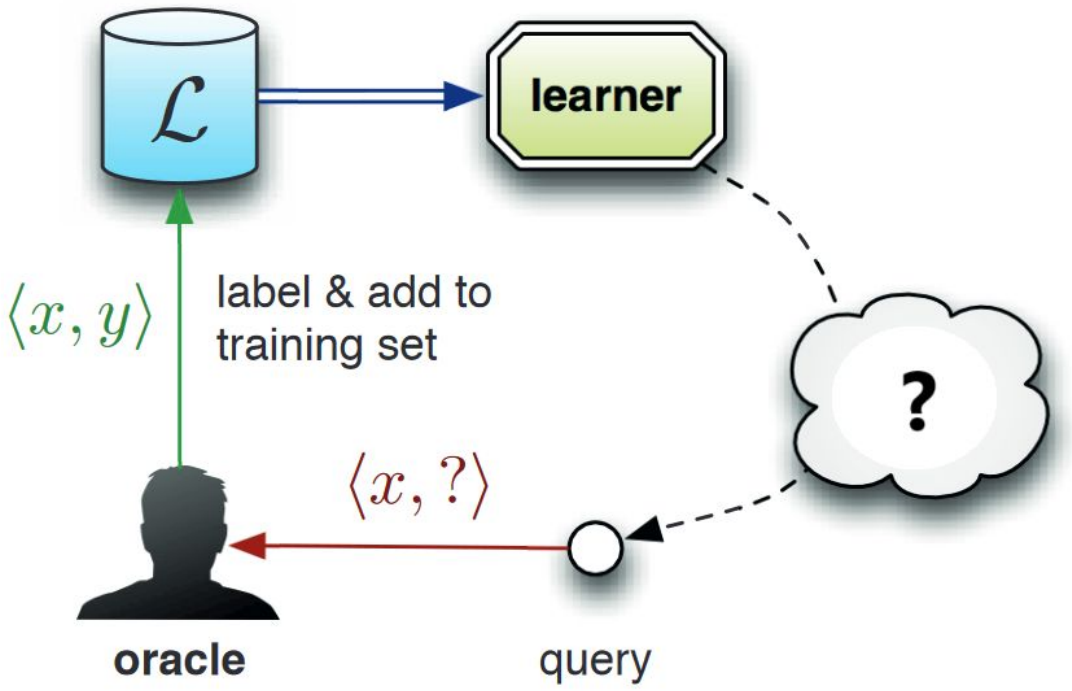
Active Learning

The Machine Learning kind

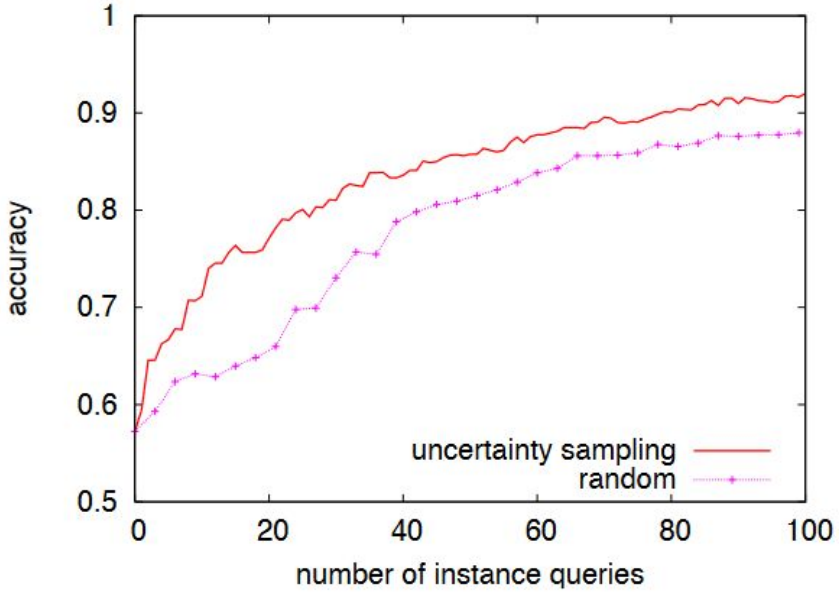
General Idea



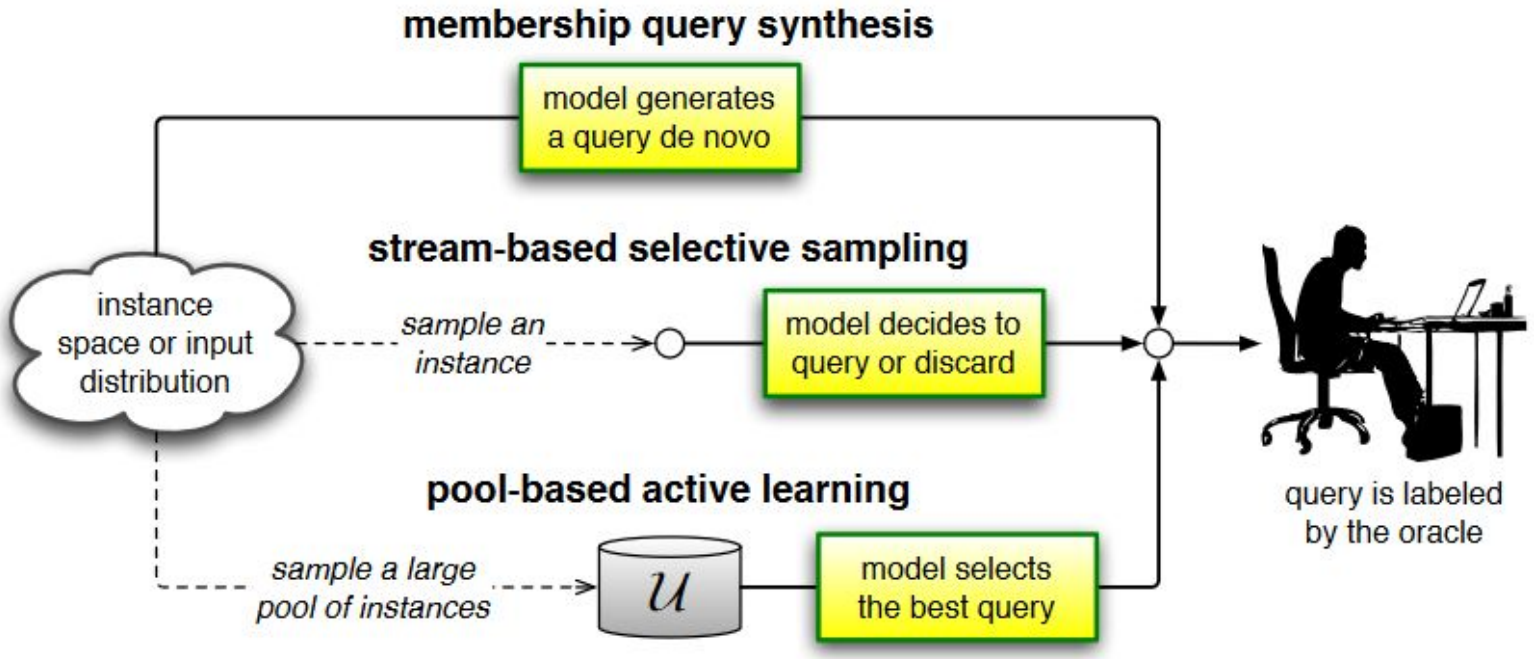
Motivation

Expensive annotation process

- Speech recognition
- Entity recognition
- Sentiment classification
- “Experimental sciences”
- DeepAL
  - Image recognition
  - Object detection



Active Learning scenarios



Query Strategy

Uncertainty Sampling

Query-By-Committee

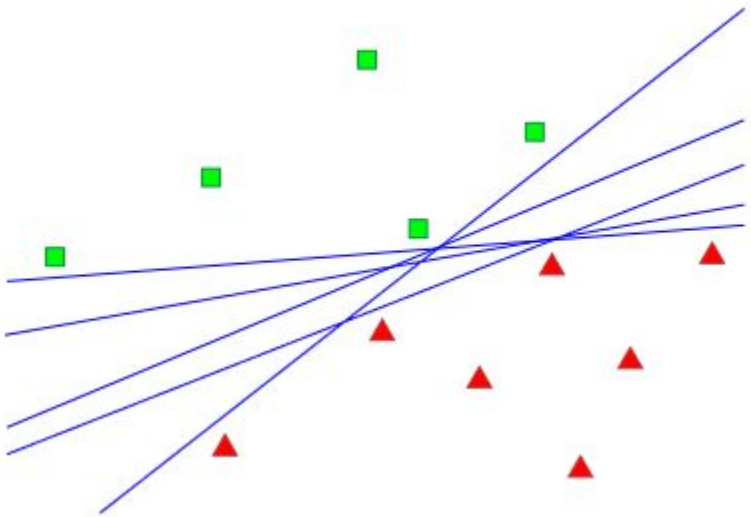
Expected Model Change

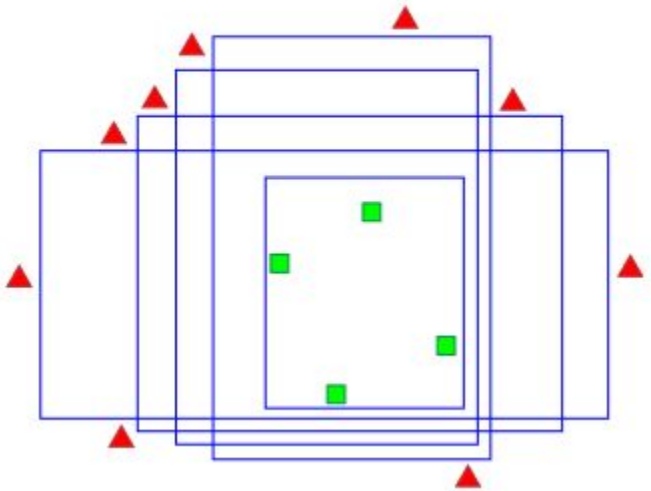
Variance Reduction

Estimated Error Reduction

(Density Weighing)

Version space





Uncertainty sampling

Needs “probability”

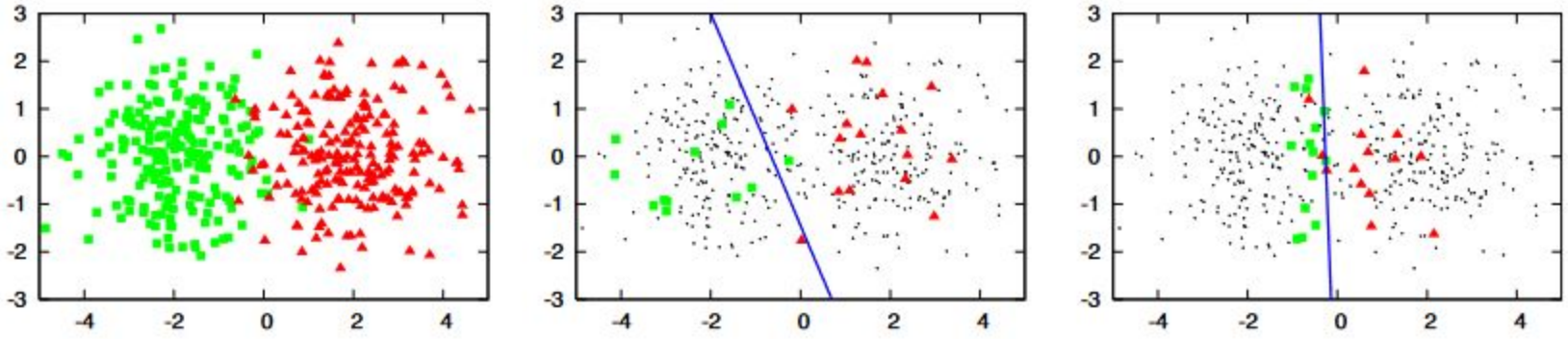
Based on Entropy

$$x^*_{ENT} = \underset{x}{\text{argmax}} - \sum_i P(y_i|x;\theta) \log P(y_i|x;\theta)$$

Least Confident labeling

$$x^*_{LC} = \underset{x}{\text{argmin}}\, P(y^*|x;\theta) \quad y^* = \text{argmax}_y\, P(y|x;\theta)$$

Uncertainty sampling + Pool-based Active learning



Query-By-Committee

A few **different** models from the version space

useful when they **disagree**

Sample model class - generative models

Use bagging/boosting

2 - 3 can be enough

Query-By-Committee

Vote entropy

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

KL - divergence

between label distributions of each committee member to the consensus

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} \sum_i P(y_i|x;\theta^{(c)}) \log \frac{P(y_i|x;\theta^{(c)})}{P(y_i|x;\mathcal{C})}.$$

Expected Model Change

Take the one that can **change the model the most**

Expected gradient length

Useful for gradient optimized models

$$x^*_{EGL} = \operatorname*{argmax}_x \sum_i P(y_i|x;\theta) \left\| \nabla \ell(\mathcal{L} \cup \langle x, y_i \rangle; \theta) \right\|$$

Error based Techniques

Reduce Error of the model, or its variance within the model class

Total error = Noise + Bias + **Variance**

Theoretically sound, impractical to compute

Further notes

Density-Weighted (Hybrid) Methods - Take samples that are more representative (~ not outliers)

$$x^*_{ID} = \underset{x}{\operatorname{argmax}}\, \phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} \operatorname{sim}(x, x^{(u)}) \right)^{\beta}$$

Batch-Mode - query a batch, simple top-N is not sufficient

AL with Costs - different queries require different amount of work - **cost**

E.g. length recording, number of tokens in text, different experiment

Stopping strategy

Hard limit in budget

Pre-defined criteria:

- # iterations
- Sought accuracy
- ...

Stabilizing Predictions

- Extra subset - see when predictions stabilize

AL & Deep Learning

Deep learning is data hungry

AL ~ study of finding the important data samples to learn with

=> Can we use DeepAL to decrease labeling cost w/o decreasing performance?

\+ Lots of unlabeled data accessible

Challenges in DeepAL

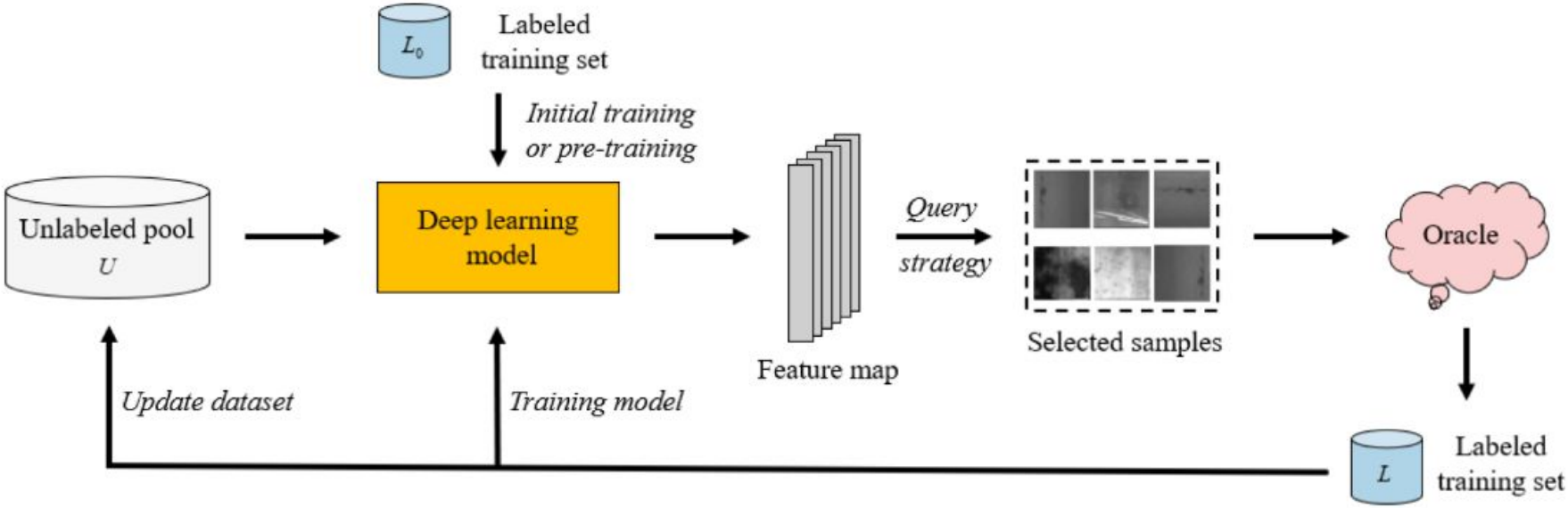
Model uncertainty - softmax can be worse than random sampling

Not enough data in a query leads to insignificant change in the model

Query strategies of AL **are incompatible** with DL

feature representations assumed *fixed in AL* but *learned in DL*

General pipeline



Query Strategy in DeepAL

Only batch mode

One-by-one querying is inefficient for DL

$$\mathcal{B}^* = \underset{\mathcal{B} \subseteq U}{\arg\max}\, a_{batch}(\mathcal{B}, f_\theta(L)),$$

Bayesian Active Learning by Disagreement - adapted to batch settings

$$a_{\text{BatchBALD}}\left(\{x_1, \ldots, x_b\}, \mathcal{P}\left(\omega \mid D_{train}\right)\right) = \mathbb{I}\left(y_1, \ldots, y_b; \omega \mid x_1, \ldots, x_b, D_{train}\right),$$

$$\mathbb{I}\left(y_{1:b}; \omega \mid x_{1:b}, D_{train}\right) = \mathbb{H}\left(y_{1:b} \mid x_{1:b}, D_{train}\right) - \mathbb{E}_{\mathcal{P}(\omega \mid D_{train})}\mathbb{H}\left(y_{1:b} \mid x_{1:b}, \omega, D_{train}\right),$$

Query Strategy in DeepAL

Uncertainty-based

Very common, despite low quality uncertainty estimates

Hybrid

Consider similarity as well

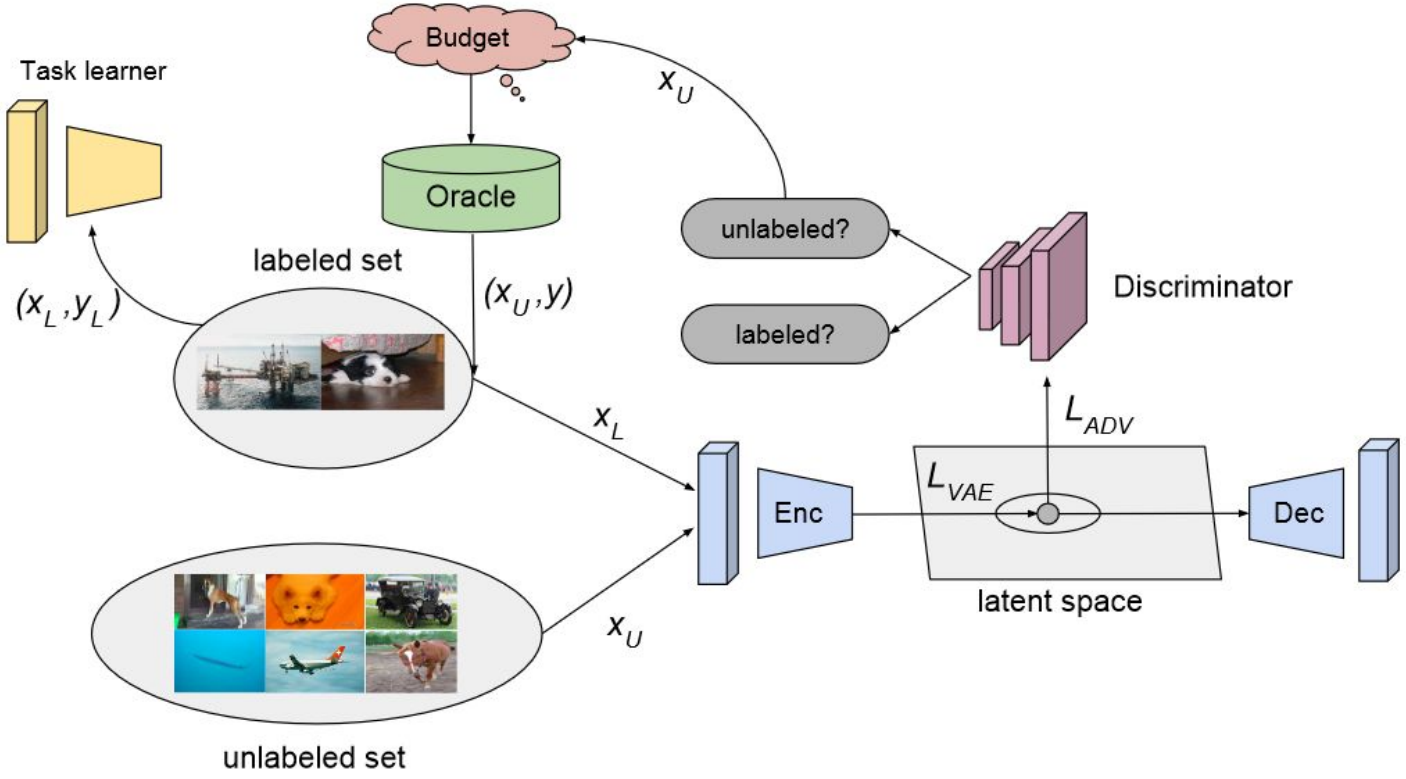
Explicit - *Exploration-P*

$$\mathrm{I}(S) = E(S) - \frac{\alpha}{|S|} R(S),$$

Implicit - *Variational Adversarial Active Learning*

Select queries using latent space representation

Variational Adversarial Active Learning (VAAL)



Query Strategy in DeepAL

Bayesian Deep Learning

Enable better estimates on uncertainty

Deep (Ensemble) Bayesian Active Learning

Monte-carlo dropout to obtain posterior samples

Possibly ensembles of NNs

BatchBALD on top

Query Strategy in DeepAL

Density-based methods

- Disregard the classification uncertainty
- Looking for **core-set**
  - Difficult to compute pairwise distances
  - Core set found to be too big for some datasets

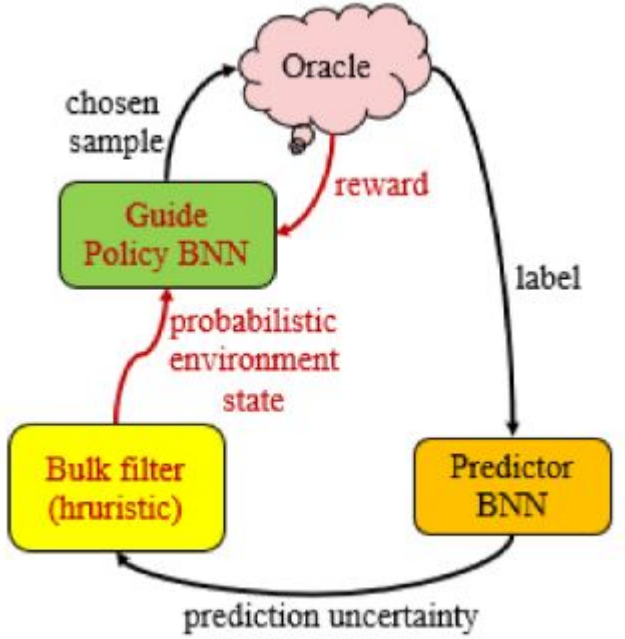
Query Strategy in DeepAL

Learning to learn

Reinforcement Learning

Train a policy that makes queries

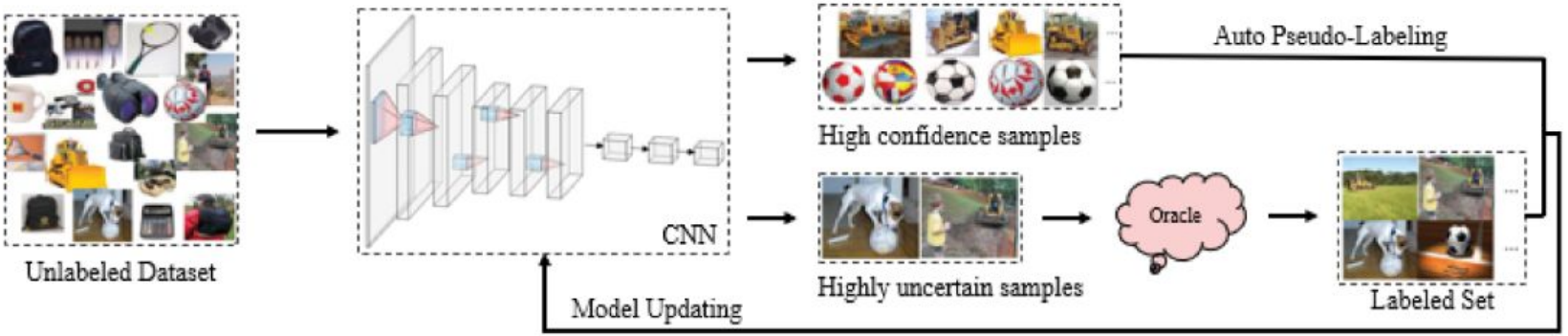
Neural Architecture Search via AL



Data expansion

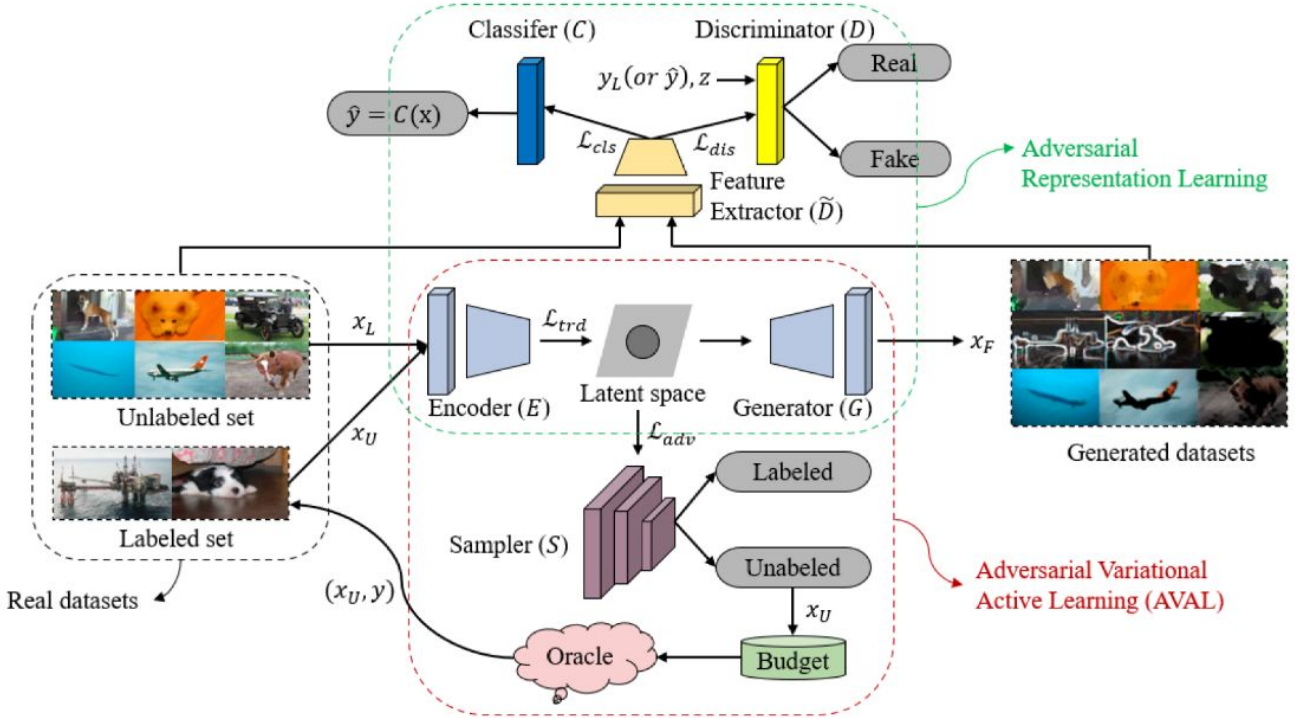
Cost-Effective AL (CEAL)

Auto-label samples with high confidence



GAN approach - *Adversarial Representation Active Learning (ARAL)*

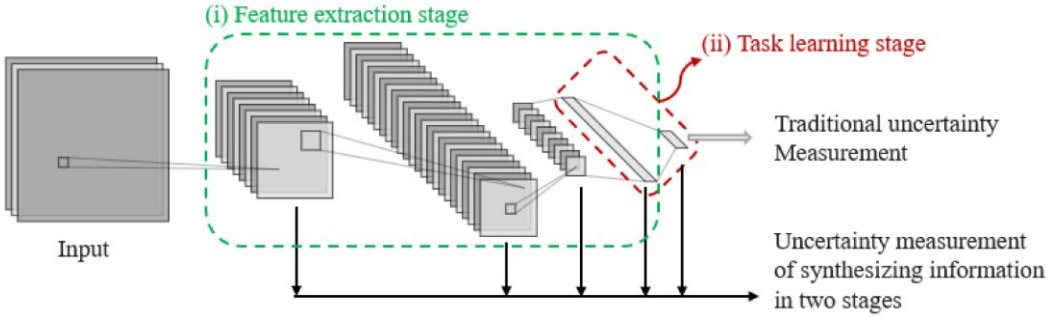
Adversarial Representation Active Learning (ARAL)



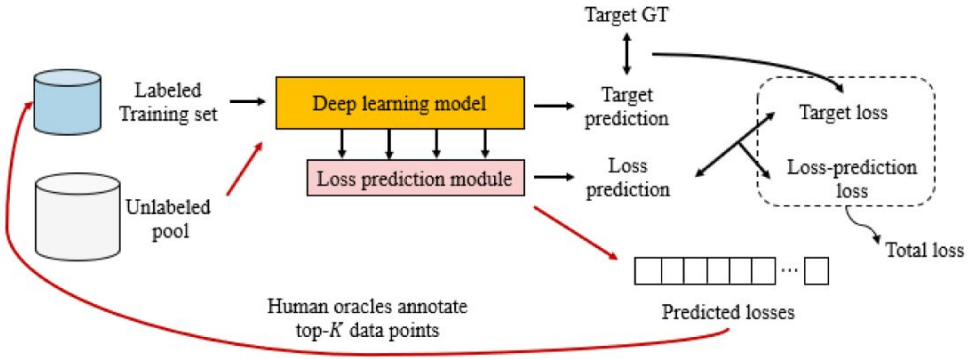
DeepAL Generic Improvements

Measure uncertainty with

more information



Learning Loss for Active Learning



DeepAL research issues

Random baseline results widely **different**

- E.g., difference by 13% on CIFAR10

General **disagreement** in what methods perform better

Models usually retrained **from scratch**

- Incremental learning?

Active Learning and LLMs

AL applied to **In-context learning**

LLMs as oracles in AL

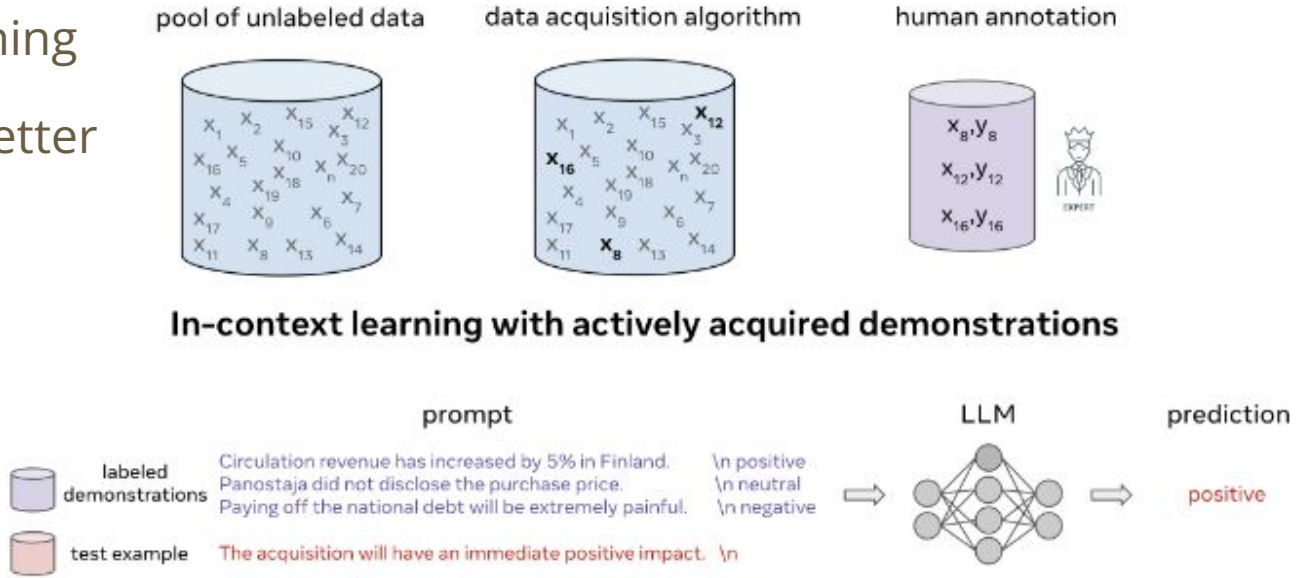
Uncertainty in LLMs

Active Learning and LLMs

In-context learning

~ Few-shot learning

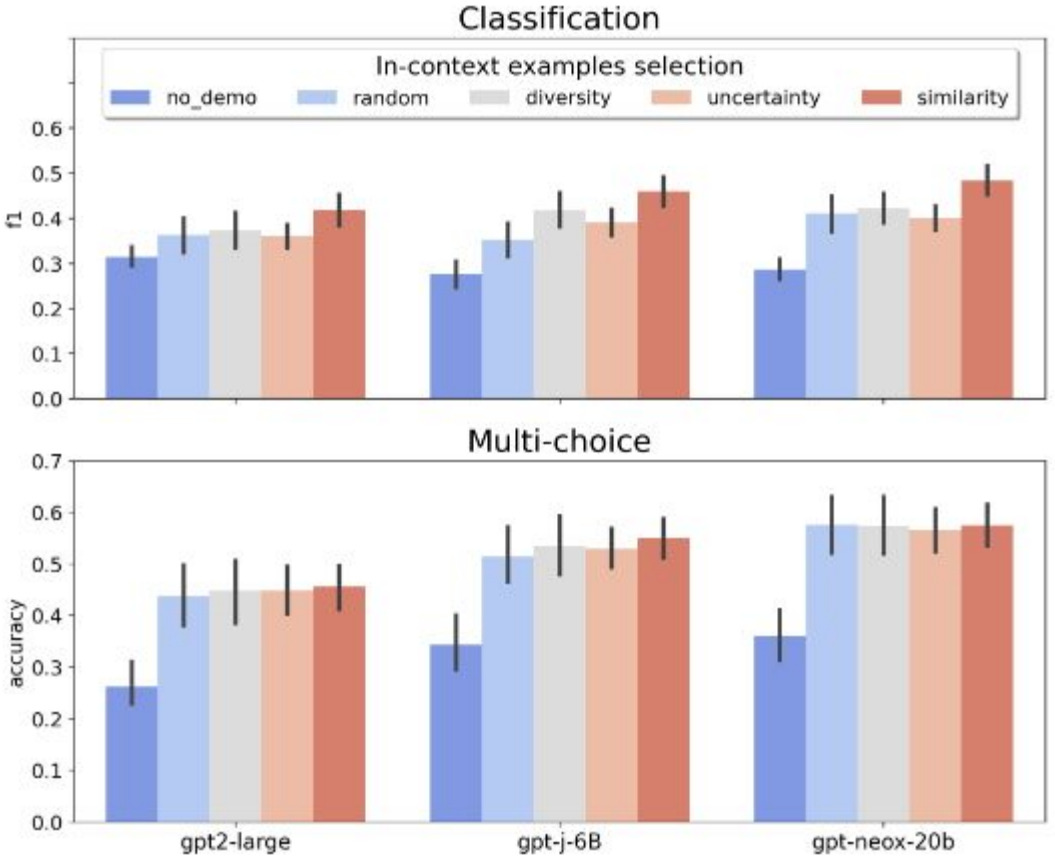
The fewer the better



In-context learning

Query strategies:

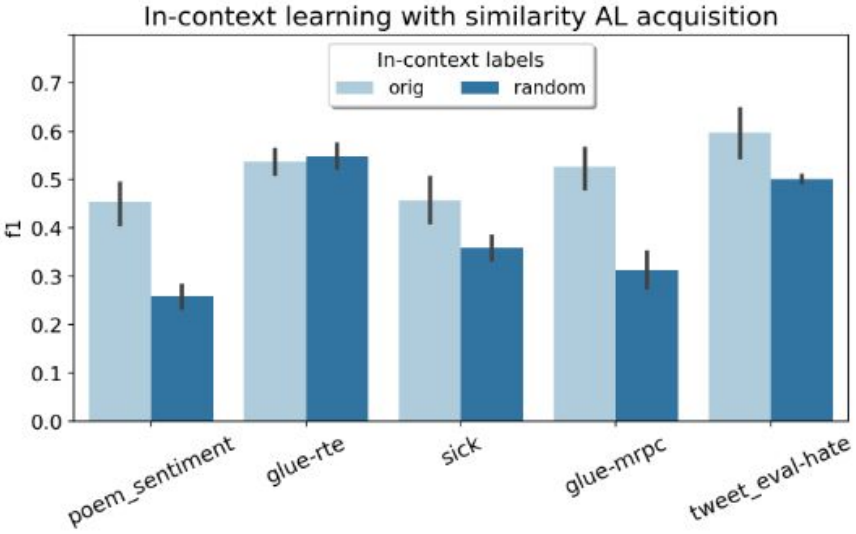
- **Diversity**
- **Uncertainty**
  - Perplexity-based
- **Similarity**
  - To the test sample

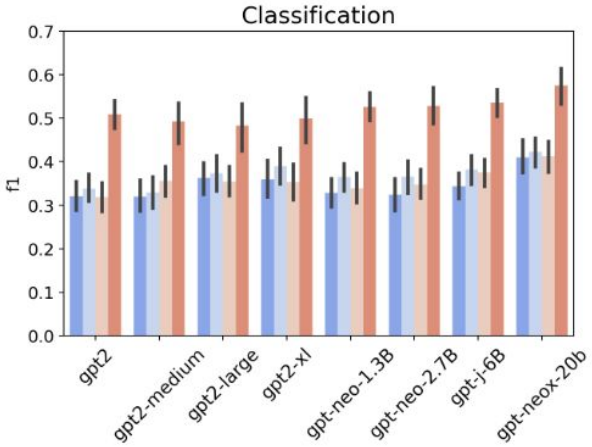


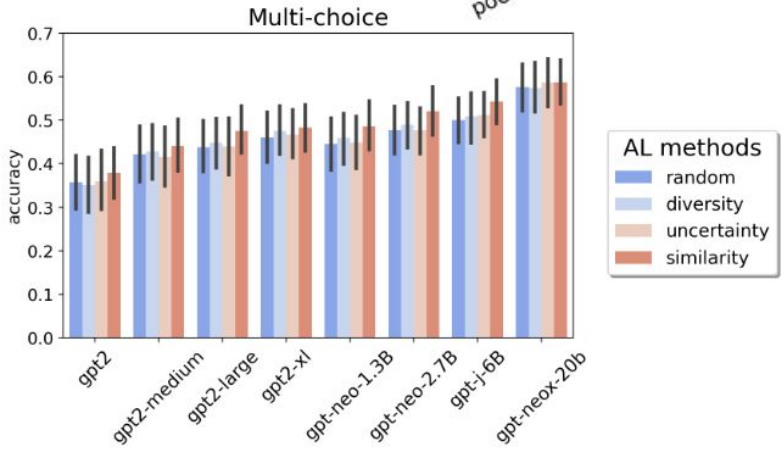
In-context learning

CAUTION: *PERSONAL OPINION*

Similarity strategy - maybe a parrot?



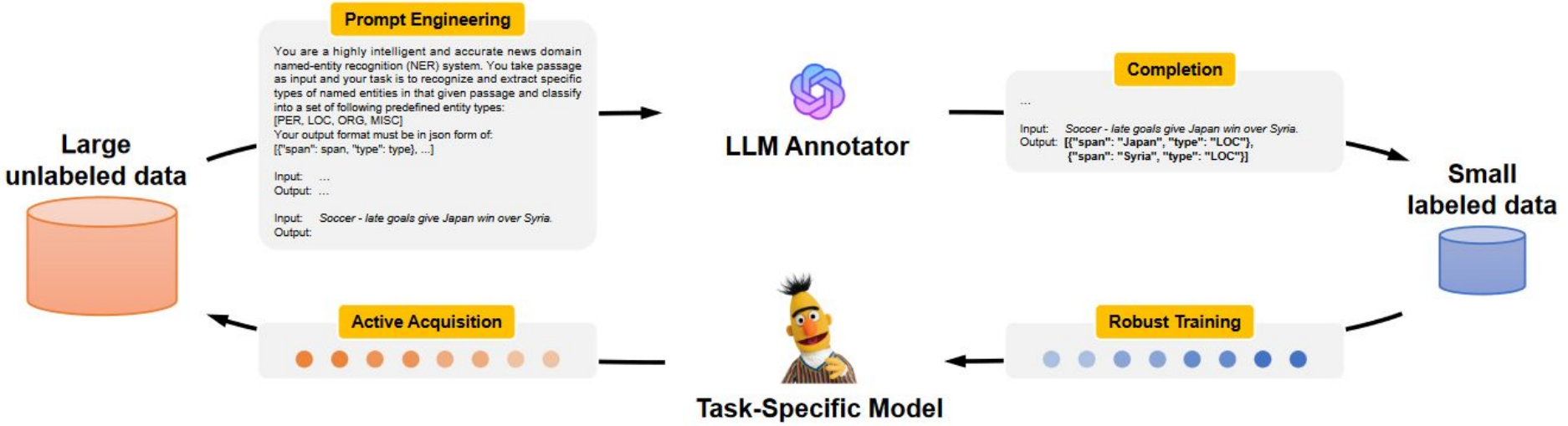




Active Learning and LLMs

LLM as an Oracle

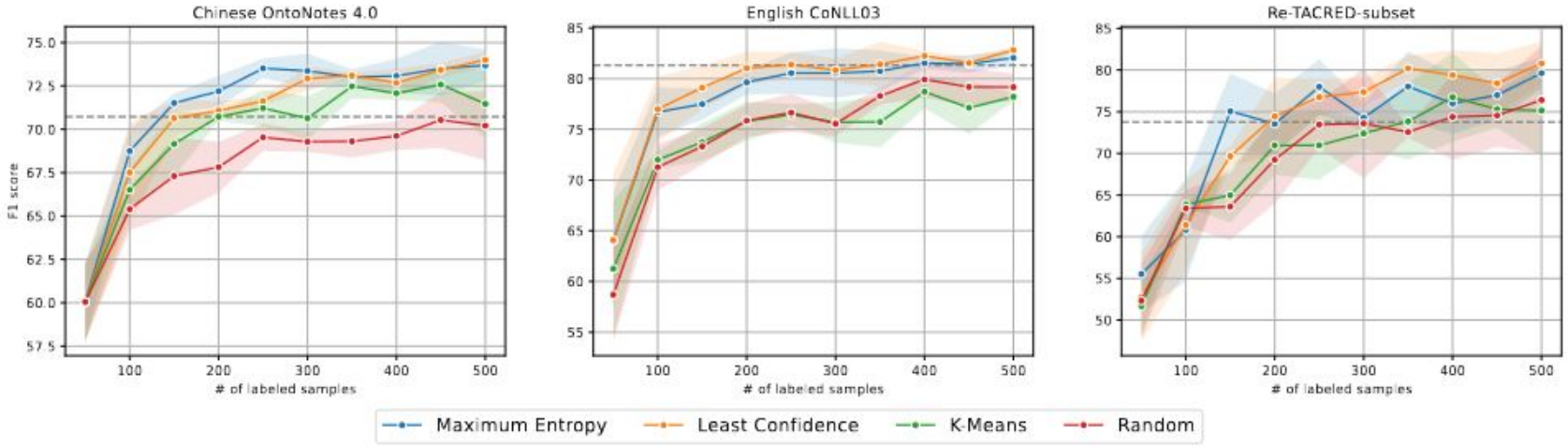
LLMaAA - LLMs as Active Annotators



LLMaAA

PROMPTING - ask the LLM directly

Backbone	Method	P	R	F1
GPT-3	PROMPTING	41.82	22.77	29.49
	LLMAAA-confidence	**57.26**	**56.09**	**56.63**
ChatGPT	PROMPTING	67.72	74.02	70.73
	LLMAAA-confidence	**72.66**	**75.49**	**74.00**
GPT-4	PROMPTING	68.70	**79.42**	73.68
	LLMAAA-confidence	**73.47**	76.42	**74.90**

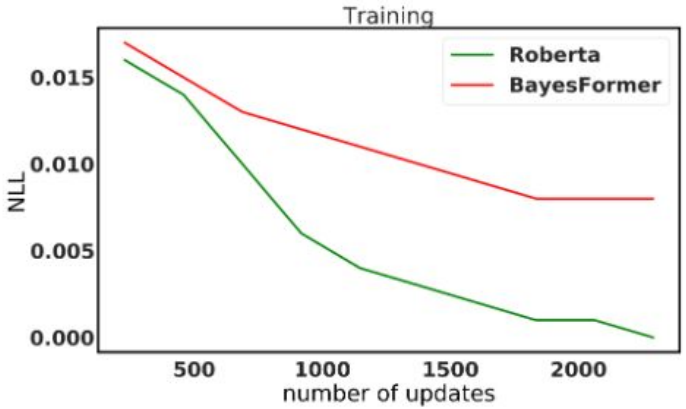


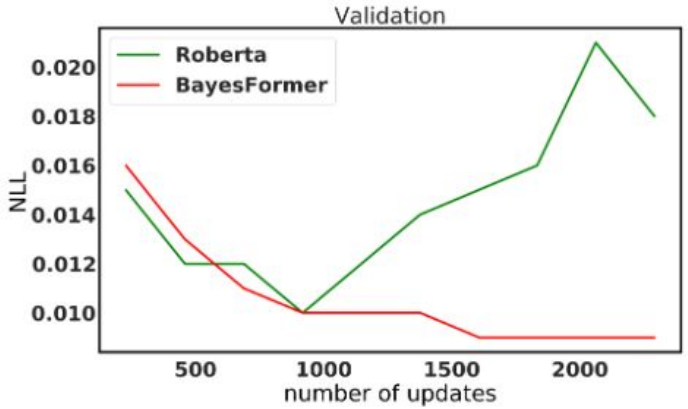
Uncertainty in LLMs

Useful in fields beyond AL

Avoiding overfitting, improving robustness

Exploration-exploitation balance in recommender systems





BayesFormer

Uses **Approximate variational inference**

Ideal:

$$p(\mathbf{y} \mid \mathbf{x}) = \int p(\mathbf{y} \mid \mathbf{x}, \mathbf{W}) \cdot p(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}) \cdot d\mathbf{W}$$

Approximation:

$$\min_q - \sum_{i=1}^{n} \int q(\mathbf{W}) \cdot \log[p(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{W})] \cdot d\mathbf{W} + \mathrm{KL}(q(\mathbf{W}) || p(\mathbf{W}))$$

Unbiased estimate of an approximation:

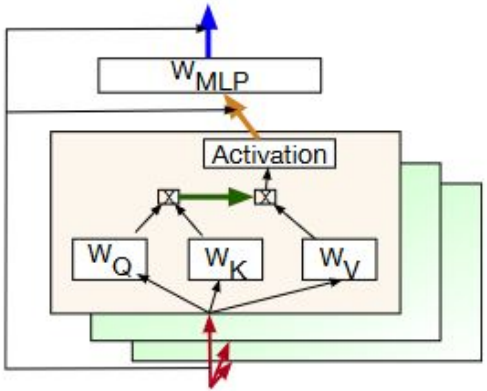
$$\min_q - \sum_{i=1}^{n} \log[p(\mathbf{y}_i \mid \mathbf{f}_{\mathbf{y}_i, \widehat{\mathbf{W}}_i}(\mathbf{X}_i))] + \mathrm{KL}(q(\mathbf{W}) || p(\mathbf{W}))$$

BayesFormer

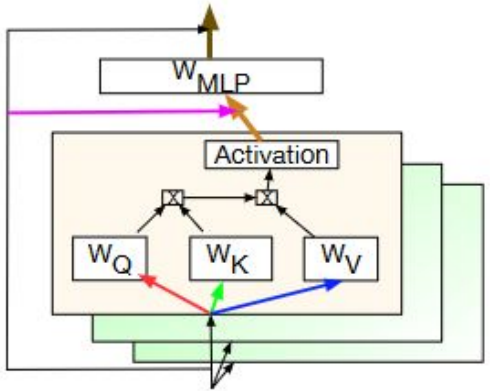
Distribution q is modeled using matrices M:

$$\mathbf{W}_{*,i} \sim p\mathcal{N}(0, \sigma^2\mathbf{I}) + (1-p)\mathcal{N}(\mathbf{M}_{*,i}, \sigma^2\mathbf{I})$$

Estimate using dropout



Classical



BayesFormer

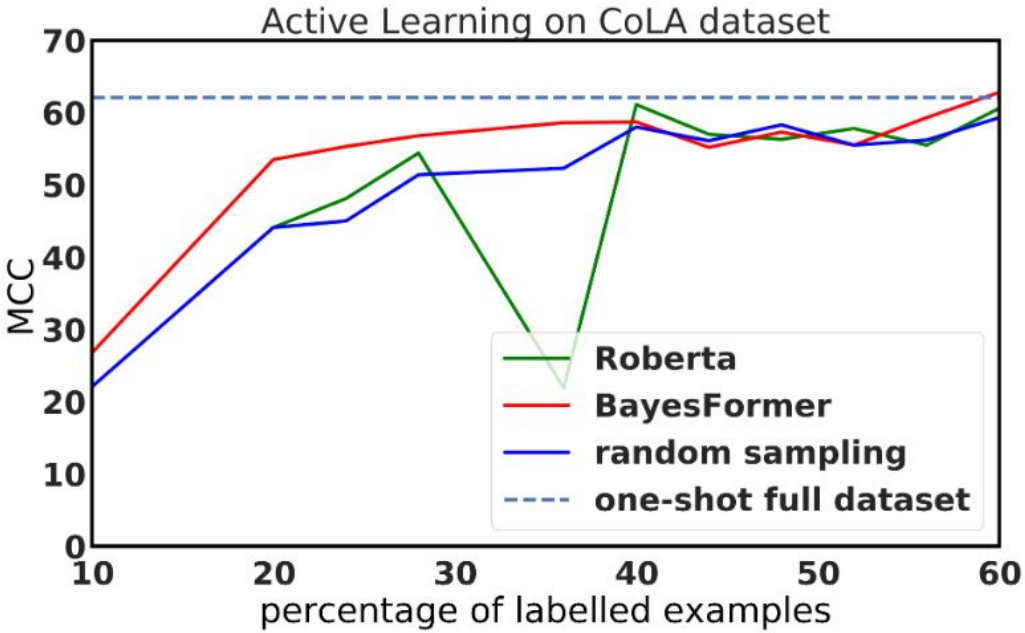
BayesFormer

Active Learning scenario:

Monte Carlo dropout

Bayesian AL with Disagreement

=> MC-BALD



References

Settles, Burr. “Active Learning Literature Survey.” Technical Report. University of Wisconsin-Madison Department of Computer Sciences, 2009. https://minds.wisconsin.edu/handle/1793/60660.

Settles, Burr. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing, 2012. https://doi.org/10.1007/978-3-031-01560-1.

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. “A Survey of Deep Active Learning.” *ACM Computing Surveys* 54, no. 9 (December 31, 2022): 1–40. https://doi.org/10.1145/3472291.

Sinha, Samarth, Sayna Ebrahimi, and Trevor Darrell. “Variational Adversarial Active Learning.” arXiv, October 28, 2019. http://arxiv.org/abs/1904.00370.

Margatina, Katerina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. “Active Learning Principles for In-Context Learning with Large Language Models.” arXiv, May 23, 2023. http://arxiv.org/abs/2305.14264.

Zhang, Ruoyu, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. “LLMaAA: Making Large Language Models as Active Annotators.” arXiv, October 31, 2023. http://arxiv.org/abs/2310.19596.

Sankararaman, Karthik Abinav, Sinong Wang, and Han Fang. “BayesFormer: Transformer with Uncertainty Estimation.” arXiv, June 1, 2022. http://arxiv.org/abs/2206.00826.