Active Learning

Carel van Niekerk

16.04.2021

The Problem

Heinrich Heine
Universität Düsseldorf

- We have a problem in which obtaining **quality labelled data** is challenging and costly.
- Challenges:
  - Medical image data need to be labelled by medical **experts**.
  - Dialogues can be **ambiguous and challenging** to label.
  - **Biases** of labellers can teach the model biases.
- To avoid these challenges we want **expert and meticulous** data labellers which may be hard to find and expensive.

- Can the model inform us **which points** would be most beneficial to label?

Outline

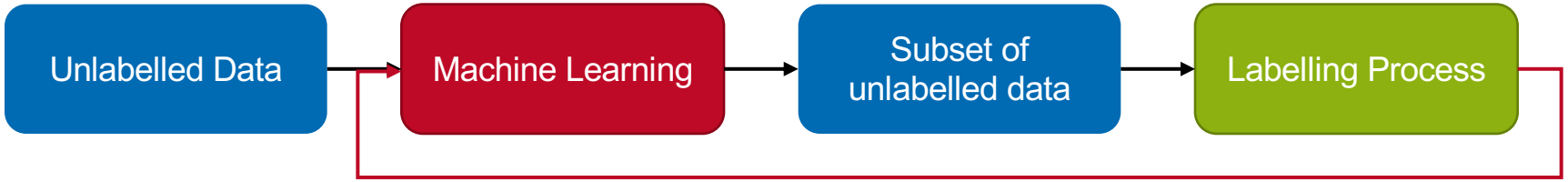
Heinrich Heine
Universität Düsseldorf

- Active Learning
  - Uncertainty
  - Disagreement
  - Bayesian
- Applications in dialogue state tracking
- Conclusion

What is Active Learning

- Supervised Learning: Learning a mapping of an input to an output based on example input-output pairs.
- Passive Learning: A set of examples is passively selected to learn from.

- Active Learning: The learner seeks to select the most useful examples to learn from based on the inputs alone.



Active Learning and Deep Learning

Heinrich Heine
Universität Düsseldorf

- Challenges:
  - Deep learning models typically need **large pools of training data**, whereas active learning requires training using small pools of data.
  - Active learning approaches tend to struggle in **high dimensional spaces**.
  - Predicting the **uncertainty** of deep learning models is not straightforward.

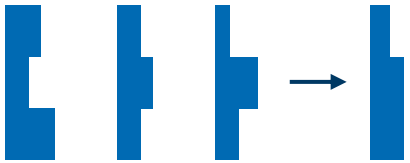
- Approaches:
  - Using the **output space** rather than parameter space for active learning reduces the dimension.
  - Techniques such as **Bayes-by-backprop, dropout and ensembles** provides more accurate uncertainty estimates for deep learning models.

Model Uncertainty

Heinrich Heine
Universität Düsseldorf

- In-domain uncertainty:
  - Various plausible models are all certain about the outcome of an in-domain point. However most of these models disagree and hence the combined predictive distribution indicates high uncertainty.

  - Various plausible models are all uncertain about the outcome of an in-domain point. Combined the predictive distribution indicates high uncertainty.



- Out-of-domain uncertainty
  - Due to point being out-of-domain plausible models should all be uncertain. Hence the combined predictive distribution would indicate high uncertainty.

Estimating Uncertainty

Heinrich Heine
Universität Düsseldorf

- Loss functions such as label smoothing can be used to produce well calibrated single model predictions.

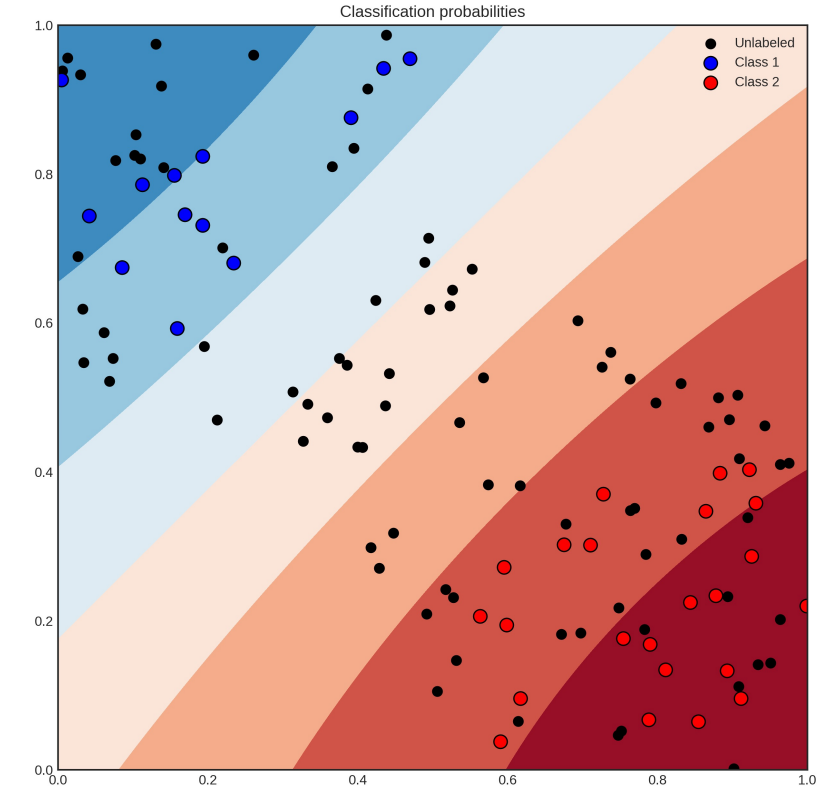
$$p(y = c|\boldsymbol{x}, \boldsymbol{D}_{train}) = \mathrm{E}_{p(\boldsymbol{w}|\boldsymbol{D}_{train})}[p(y = c|\boldsymbol{x}, \boldsymbol{w})]$$
$$\approx p(y = c|\boldsymbol{x}, \hat{\boldsymbol{w}})$$

- Utilise bayes-by-backprop, Monte Carlo dropout or an ensemble of models to obtain a set of plausible models $\left(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(M)}\right)$.

$$p(y = c|\boldsymbol{x}, \boldsymbol{D}_{train}) = \mathrm{E}_{p(\boldsymbol{w}|\boldsymbol{D}_{train})}[p(y = c|\boldsymbol{x}, \boldsymbol{w})]$$
$$\approx \frac{1}{M}\sum_{m=1}^{M} p\left(y = c|\boldsymbol{x}, \boldsymbol{w}^{(m)}\right)$$

General Approach

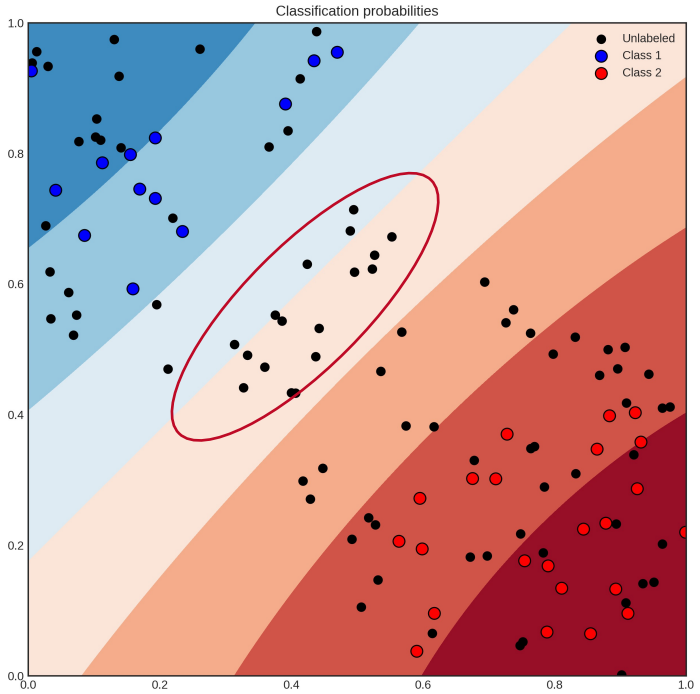
1. Create the **seed dataset**. This is a small labelled dataset used to pretrain the model.
2. Train model F_0 using the seed.
3. Collect a pool of unlabelled points $x_1, x_2, \ldots, x_N$.
4. For each input point calculate the value of the **acquisition function**, using model F_{i-1}.
5. **Select the pool of points** with the highest acquisition values.
6. **Obtain labels** for these points and **update model** to obtain F_i.
7. Repeat 4-6 until a stopping criteria is reached.



Acquisition Functions
Uncertainty

Heinrich Heine
Universität Düsseldorf

- We seek to find the **subset of the data for which the model is most uncertain** about the outcomes.

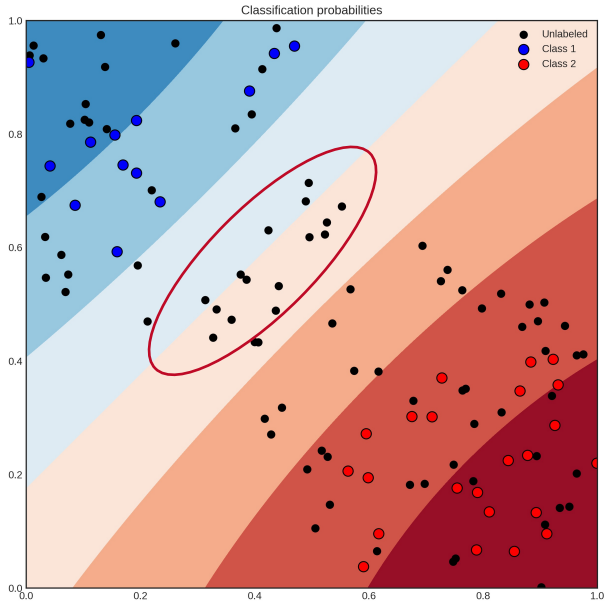


Acquisition Functions
Uncertainty - Least Confidence

Heinrich Heine
Universität Düsseldorf

- The confidence of the model can be measured by the predicted probability for the most likely outcome. The least confidence acquisition utilises this form of confidence to select points for which the model is least confident. Hence:

$$a(x) := 1 - \max_{c} p(y = c | \boldsymbol{x}, \boldsymbol{D}_{train})$$

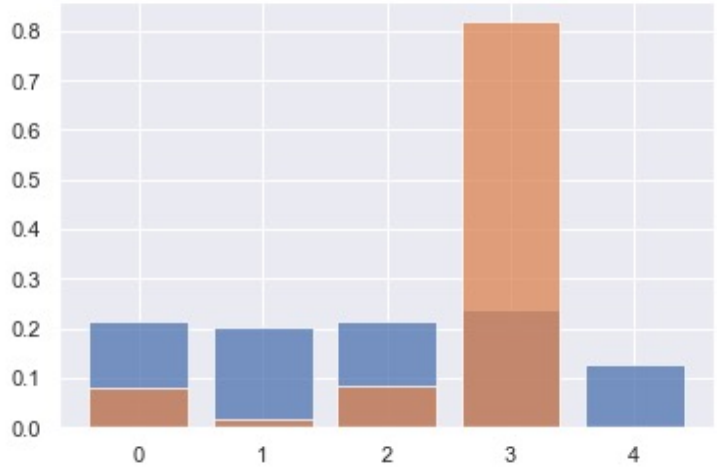


Acquisition Functions
Uncertainty - Entropy

Heinrich Heine
Universität Düsseldorf

- The entropy of a distribution is the average level of “uncertainty” present in the distribution. The Shannon entropy is defined as:

$$\mathrm{H}[p(y|\boldsymbol{x}, \boldsymbol{D}_{train})] \coloneqq -\sum_{c} p(y = c|\boldsymbol{x}, \boldsymbol{D}_{train}) \log p(y = c|\boldsymbol{x}, \boldsymbol{D}_{train})$$

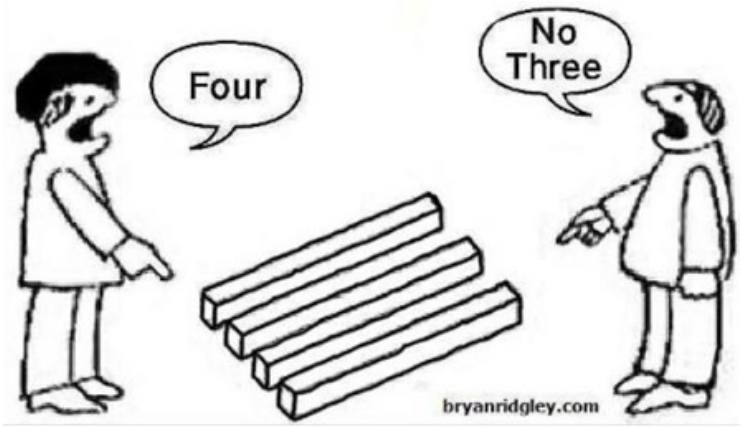


- These uncertainty metrics can indicate both in-domain uncertainty and uncertainty arising from out of domain data. They cannot distinguish between these two scenarios.

Acquisition Functions
Disagreement

Heinrich Heine
Universität Düsseldorf

- We seek to find the subset of the data for which there is the most disagreement among plausible models.
- Uncertainty arises from disagreement when various plausible models are all confident about the outcome, but disagree with each other. This typically is the case for challenging in-domain points.
- Disagreement measures can better distinguish between in-domain uncertainty and uncertainty arising from out of domain data.



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.*

Acquisition Functions
Disagreement - Variation Ratios

Heinrich Heine
Universität Düsseldorf

- The variation ratio is measure of dispersion within a ensemble. A low ratio indicates little dispersion (uncertainty) in the ensemble. A high ratio indicates a high level of dispersion (uncertainty) in the ensemble.

$$a(x) := 1 - \frac{f_M}{T}$$

- Where f_m is the number of ensemble members which predicts the modal class in the ensemble. The ensemble consist of T members.
- This measure does not consider the uncertainty of the individual models, only the uncertainty arising from disagreement.

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*.

Acquisition Functions
Mean Standard Deviation

Heinrich Heine
Universität Düsseldorf

- The mean standard deviation across classes is a similar measure of dispersion.

$$a(x) \coloneqq \frac{1}{C}\sum_{c} \mathrm{std}_{p(\boldsymbol{w}|\boldsymbol{D}_{train})}[p(y = c|\boldsymbol{x}, \boldsymbol{D}_{train})]$$

- Typically used with ensemble models to estimate the level of dispersion in the ensemble.

Acquisition Functions
Bayesian Active Learning by Disagreement

Heinrich Heine Universität Düsseldorf

- We seek to find points where the overall uncertainty is high, but the individual model uncertainty is low. Hence the models disagree on the outcomes.

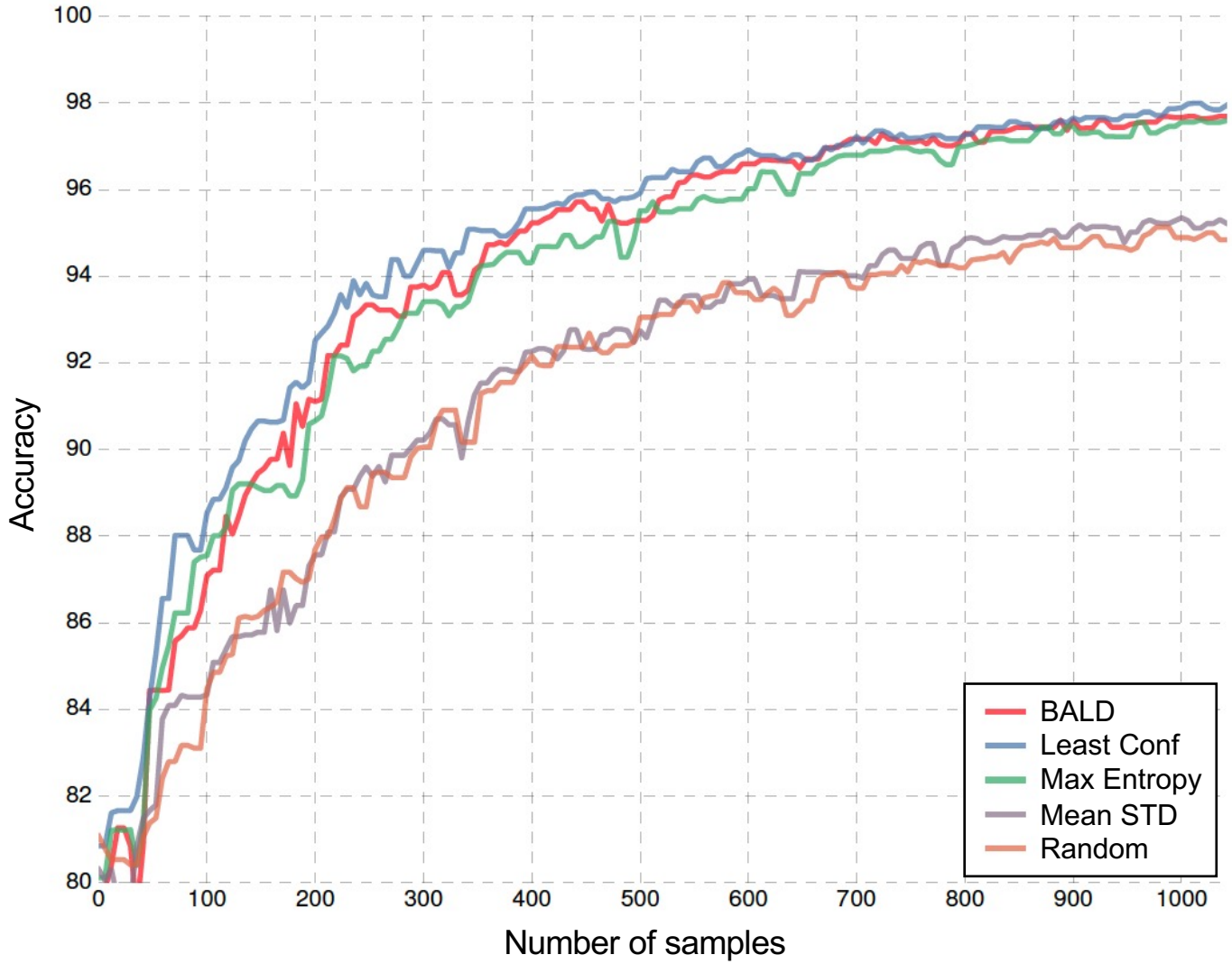
$$a(x) = I[y, \boldsymbol{w} | \boldsymbol{x}, \boldsymbol{D}_{train}] = \mathrm{H}[y | \boldsymbol{x}, \boldsymbol{D}_{train}] - \mathrm{E}_{p(\boldsymbol{w}|\boldsymbol{D}_{train})}[\mathrm{H}[y | \boldsymbol{x}, \boldsymbol{w}]]$$

$$\mathrm{E}_{p(\boldsymbol{w}|\boldsymbol{D}_{train})}[\mathrm{H}[p(y | \boldsymbol{x}, \boldsymbol{w})]] \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{H}[p(y | \boldsymbol{x}, \boldsymbol{w}^{(m)})]$$

- A point would have a high acquisition value if **individual models are certain** about its outcome, but overall **models do not agree** on the outcome.
- Entropies are calculated in **lower dimensional outcome space** rather than in extreme high dimensional parameter space.

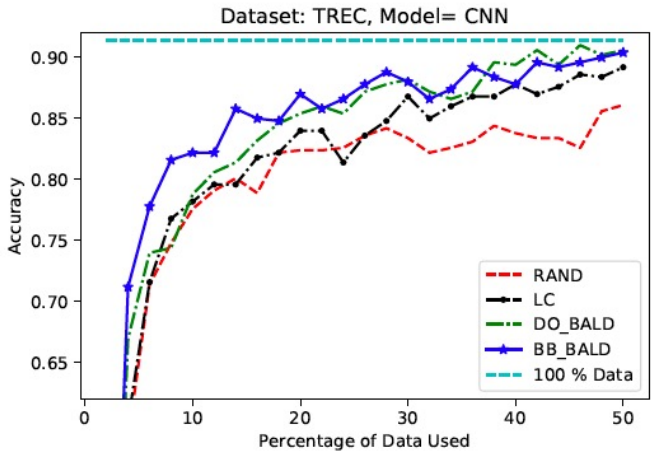
Neil Houlsby, et al. "Bayesian Active Learning for Classification and Preference Learning." (2011).

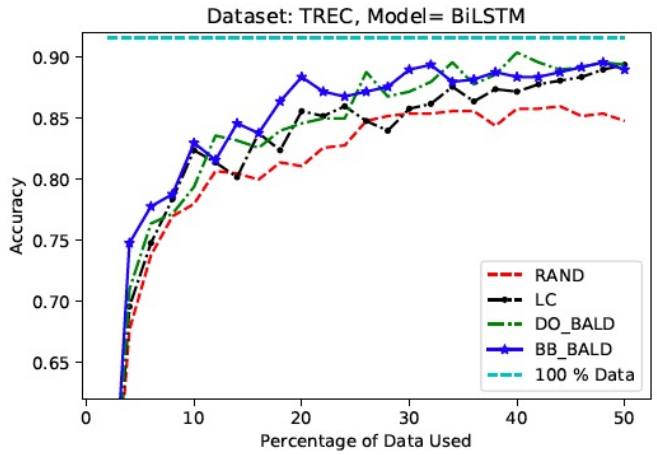
Bayesian CNN - MNIST Performance

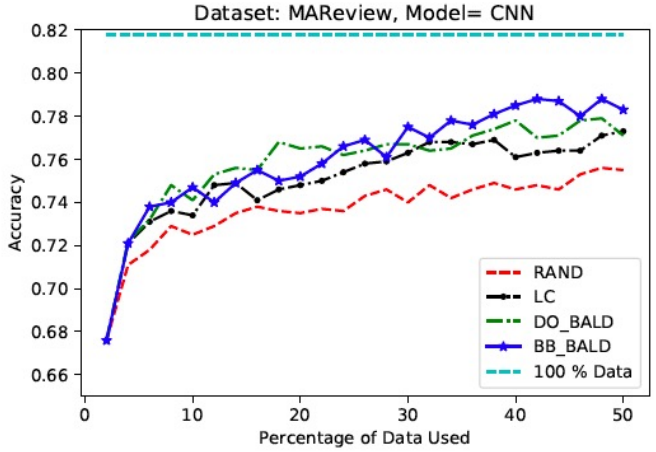


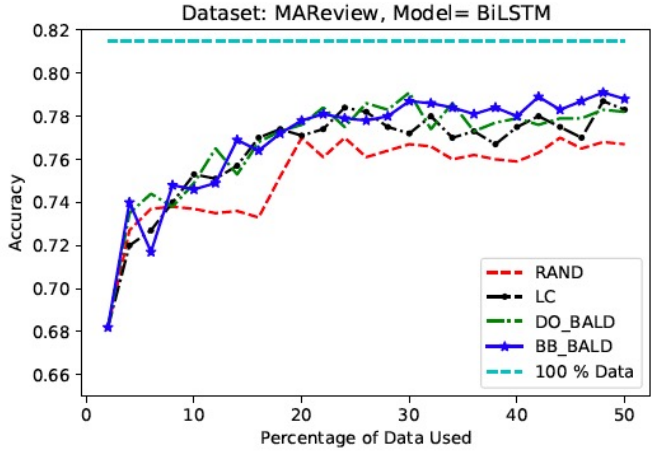
Yarin Gal and Riashat Islam and Zoubin Ghahramani, . "Deep Bayesian Active Learning with Image Data". *CoRR* abs/1703.02910. (2017).

Performance on Question Answering



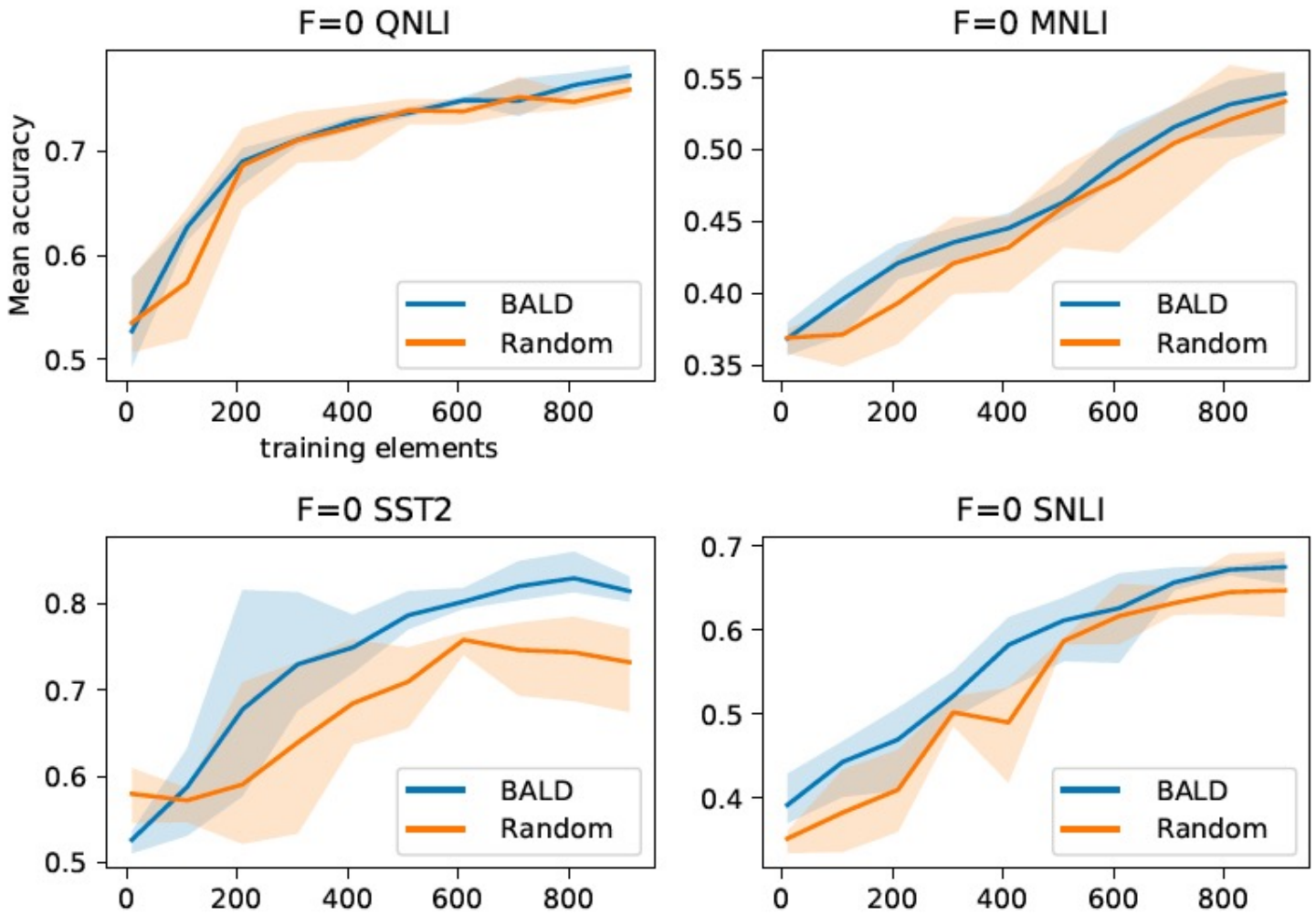






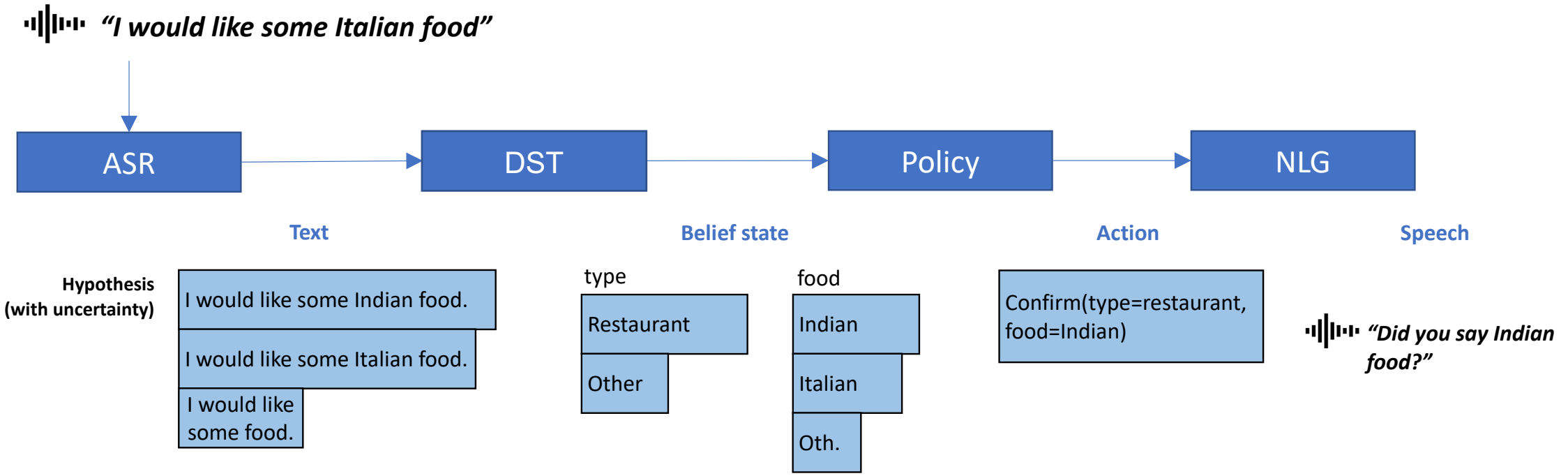
Aditya Siddhant and Zachary C. Lipton, . "Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study". *CoRR* abs/1808.05697. (2018).

Language Understanding Performance



Griesshaber, Daniel et al. "Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning". *arXiv preprint arXiv:2012.02462*. (2020).

Dialogue State Tracking



Dialog State Tracking

Heinrich Heine
Universität Düsseldorf

- Kai Xie et al - "Cost-Sensitive Active Learning for Dialogue State Tracking." illustrate the usefulness of maximum entropy active learning in DST.
- They show AL can be used for both cost effective learning on DSTC2 data.
- Next Steps:
  - Use ensemble for cost effective learning on MultiWOZ.
  - Few Shot adaptation to new domains in MultiWOZ.

Conclusion

Heinrich Heine
Universität Düsseldorf

- Pros
  - Reduces the amount of quality labelled data required to train models.
- Cons
  - Relies on the quality of labelled data.
  - It is challenging to select a good uncertainty measure. To select the optimal measures we often need the full labelled dataset.
- Potentials
  - Low resource learning for dialogue state tracking
  - Few shot adaptation of dialogue state tracking approaches