Evaluation of Active Learning Strategies for Transformer Architecture based Language Models

Ada Güney Arslan

Introduction

Problem Definition and Solution Proposal

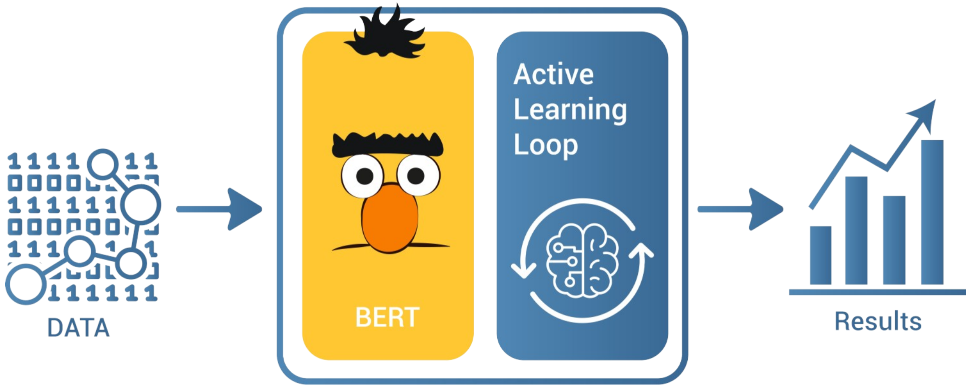
Dresden Database Systems Group

Problem

- Can active learning improve the performance of hidden transformer blocks?

Solution

- Usage of deep pre-trained transformer architecture based models
- Usage of pool based uncertainty sampling for active learning
- Performance evaluation and results



Foundations

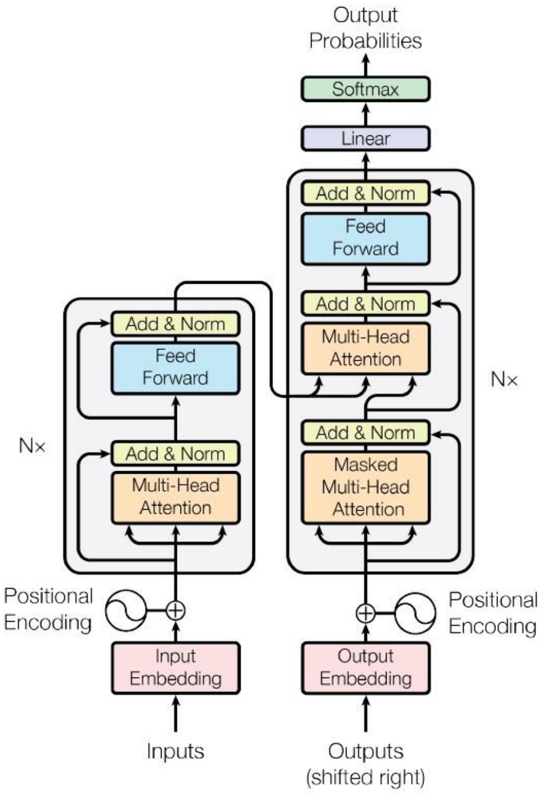
Transformers

Details on Transformers

- Proposes transformers instead of recurrence and convolutions

Attention Mechanism

- Most important part
- Bypasses sequential processing, enables parallelization
- Longer sequences



BERT Model

Dresden Database Systems Group

Details on BERT model

- Stands for Bidirectional Encoder Representations from Transformers

Pre-Training

- Cloze task
- Next sentence prediction

Active Learning

What is it?

- Iterative and supervised
- Query an information source
- Label new data points

What should we pay attention to?

- Class Imbalances
- Binomial and Multinomial data
- Costliness of data-labeling

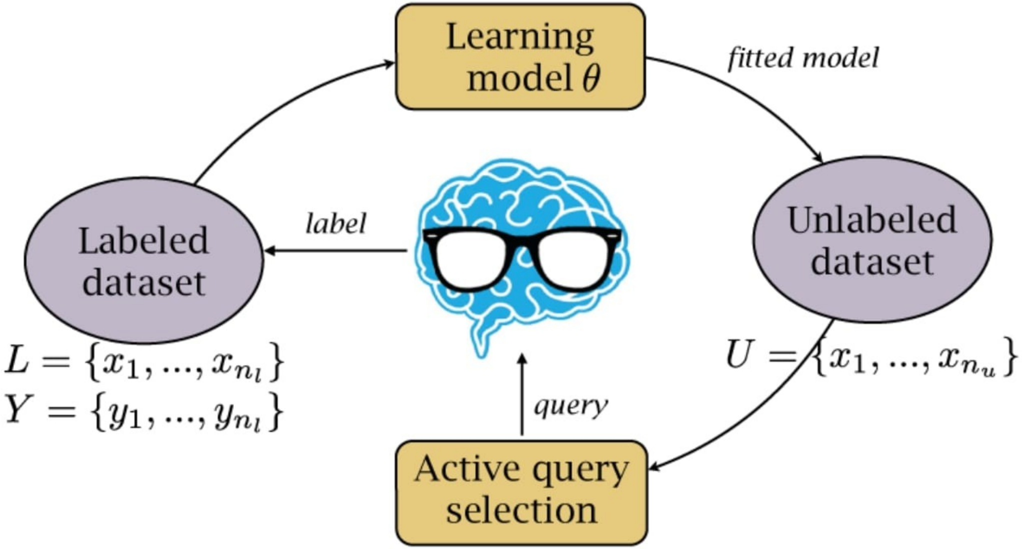


Figure:
https://deepai.org/machine-learning-glossary-and-terms/active-learning, 2022

Active Learning

Logistic regression

- Binomial and multinomial cases
- Probabilistic

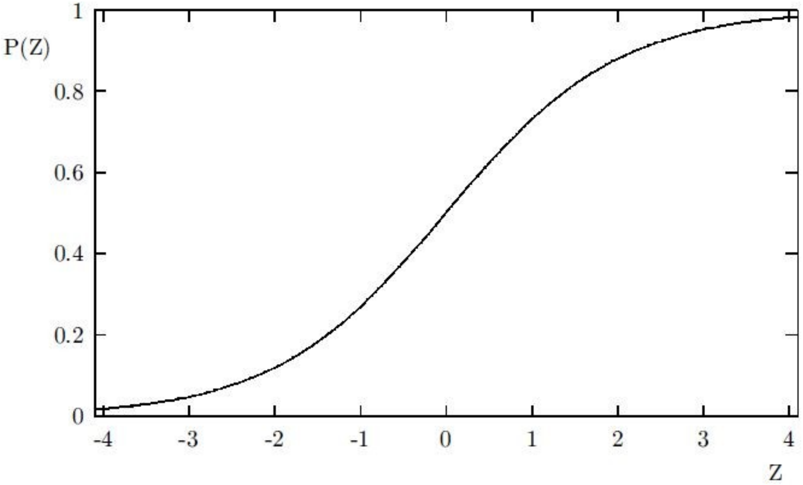


Figure: Logistic Curve P(Z) [Cramer, 2003]

Least confidence

- Batch of least confident samples

$$\phi^{LC}(x) = 1 - P(y^* | x; \theta)$$

Formula: Least Confidence [Culotta and McCallum, 2005]

Dresden Database Systems Group

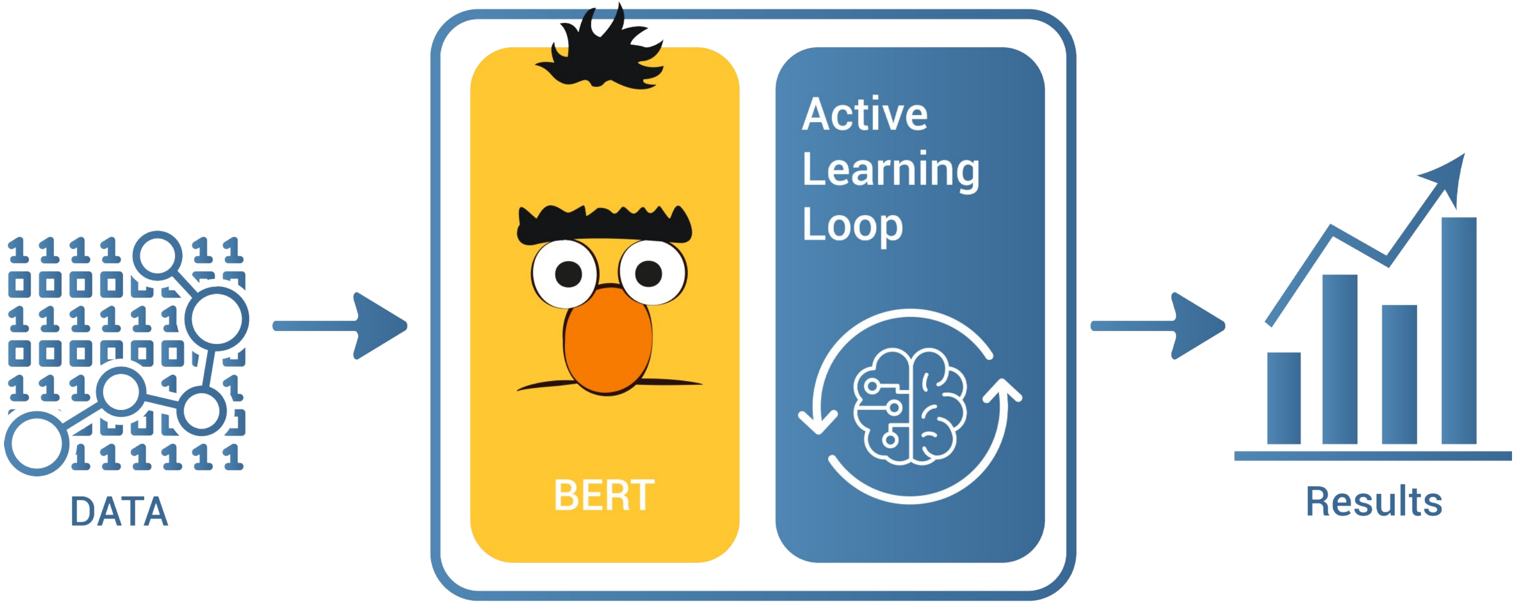
Implementations

Experiment Setup

Dresden Database Systems Group

Setup

- Pre-processing via BERT-BASE
- Pool based uncertainty sampling
- Sharing a metric as visualization

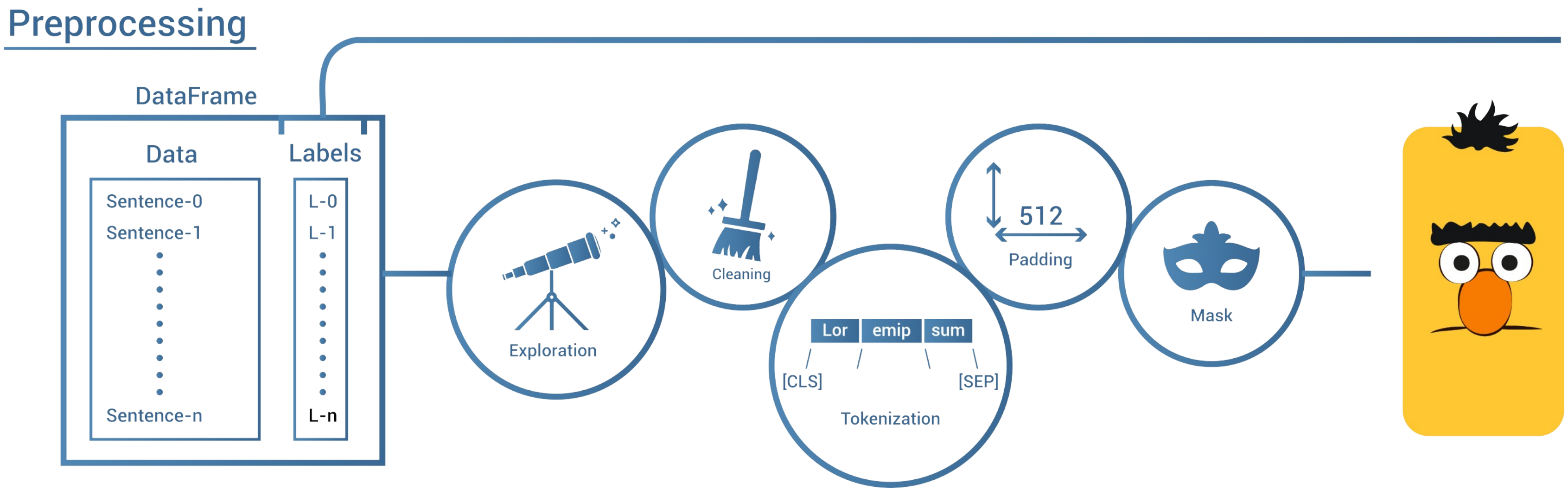


Pre-processing

Dresden Database Systems Group

Data as an Input for BERT

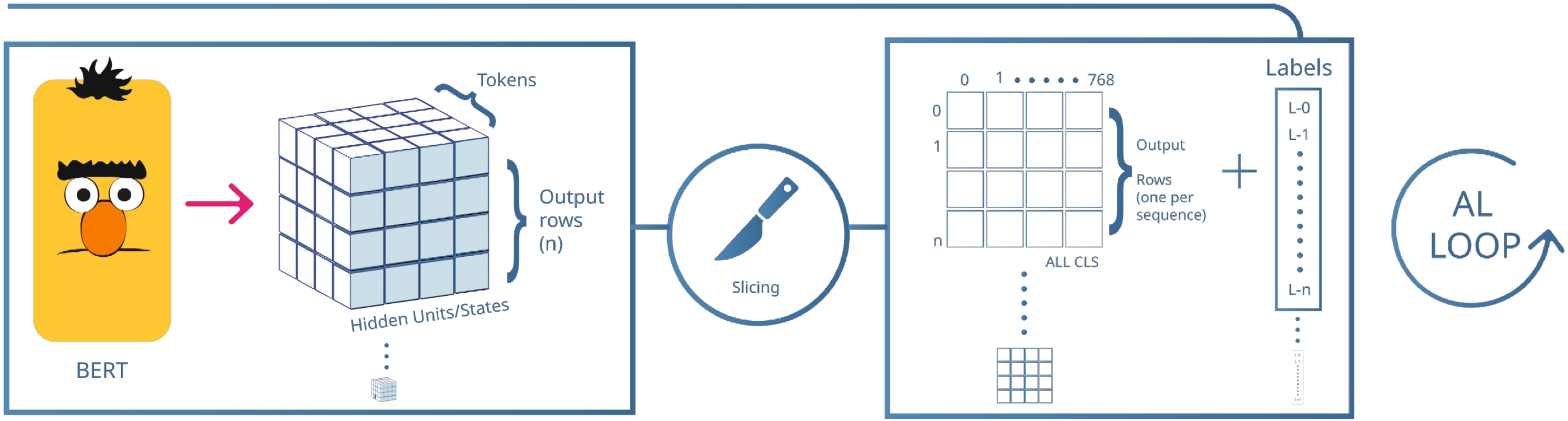
- Exploration, Cleaning, Tokenization, Padding
- Attention mask and padded as an input



Pre-processing

Extraction of Classification Tokens

- Slice the CLS tokens
- Save tensors as a matrix of features
- Features and pre-existing labes for active learning

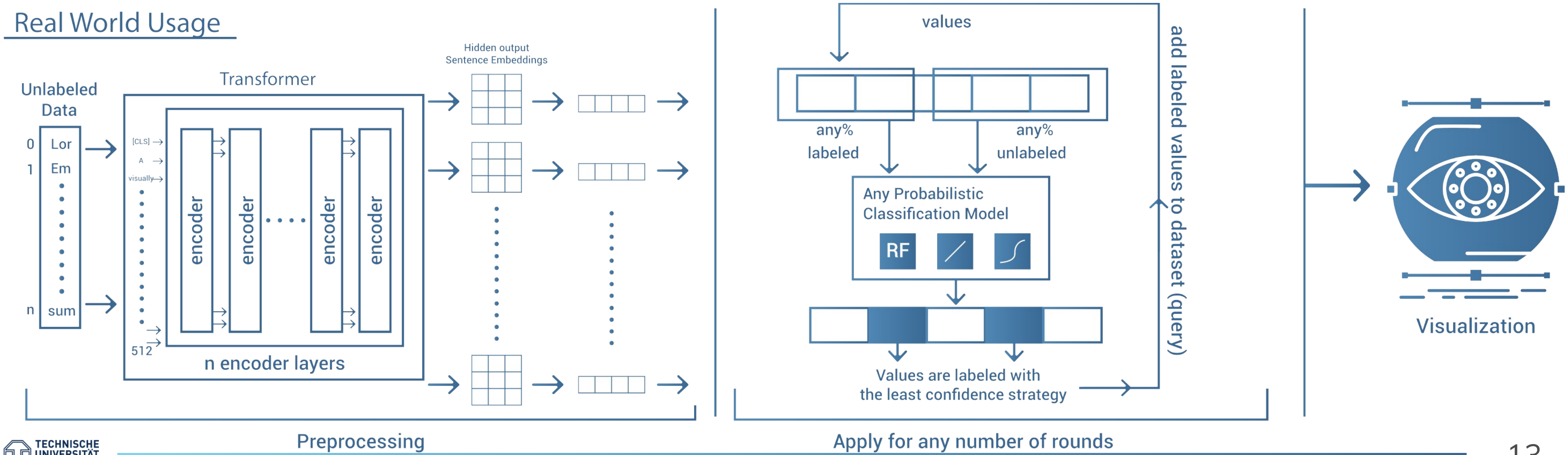


Active Learning

Dresden Database
Systems Group

Active Learning Loop

- Logit model trianed with labeled data
- Unlabeled data predicted
- Labeled with a strategy



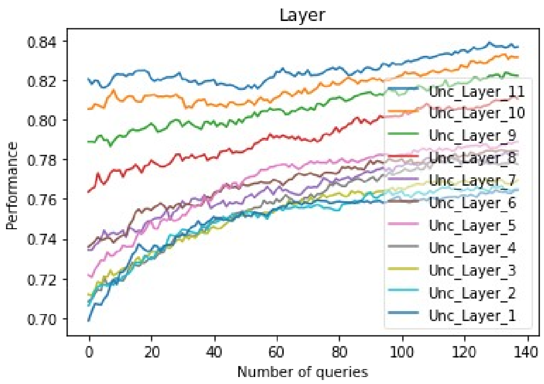
Evaluation

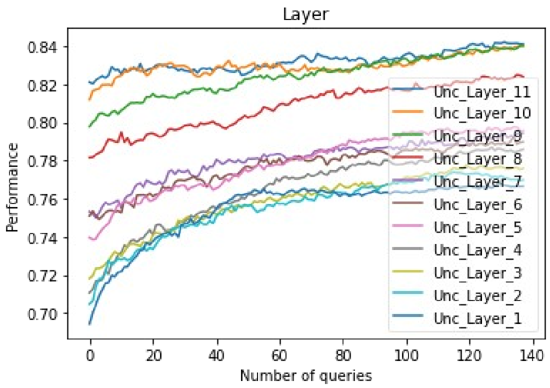
Addition vs Concatenation

Dresden Database Systems Group

Motivation

- Combining different vectors and achieving better performance
- Decision: addition





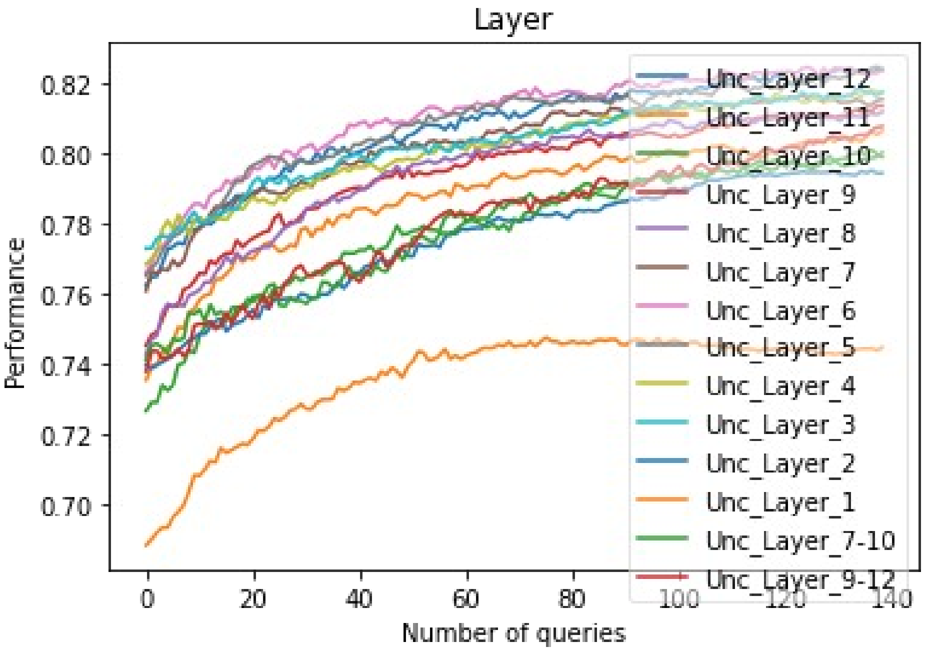
Comparison Metrics

- Length change
- Query time influence
- Informativeness

Results IMDB

Data-set

- Polar movie reviews
- Labels; negative or positive (balanced)

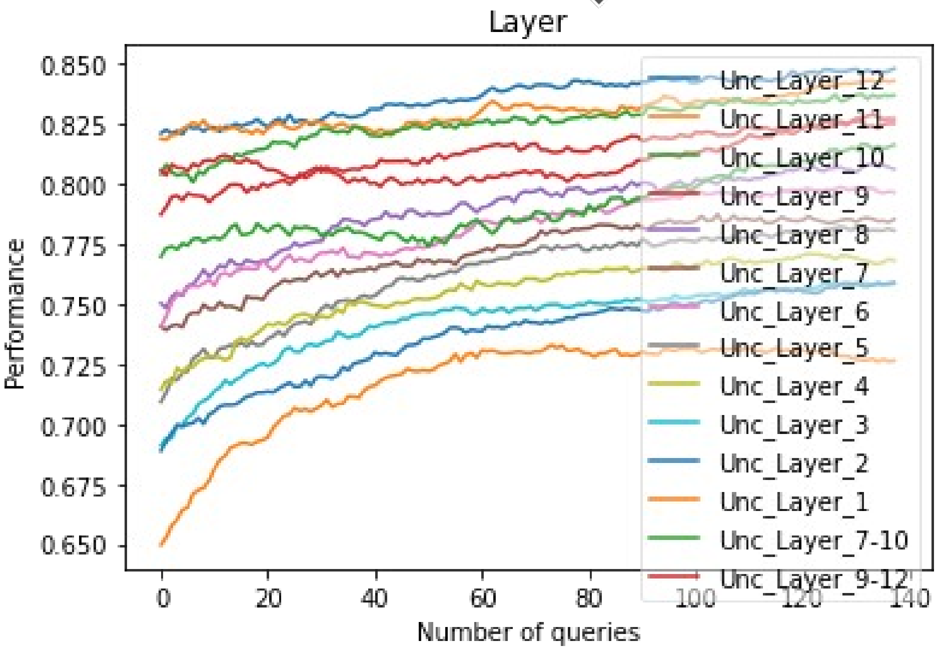


1	2	3	4	5	6	7	8	9	10	11	12	7.10	9.12
73.6	77.6	80.3	80.2	80.8	81.1	80.3	79.5	79.5	78.0	78.8	80.7	77.8	78.1

Results SST

Data-set

- Fine grained sentiment phrases
- Labels; negative or positive (balanced)

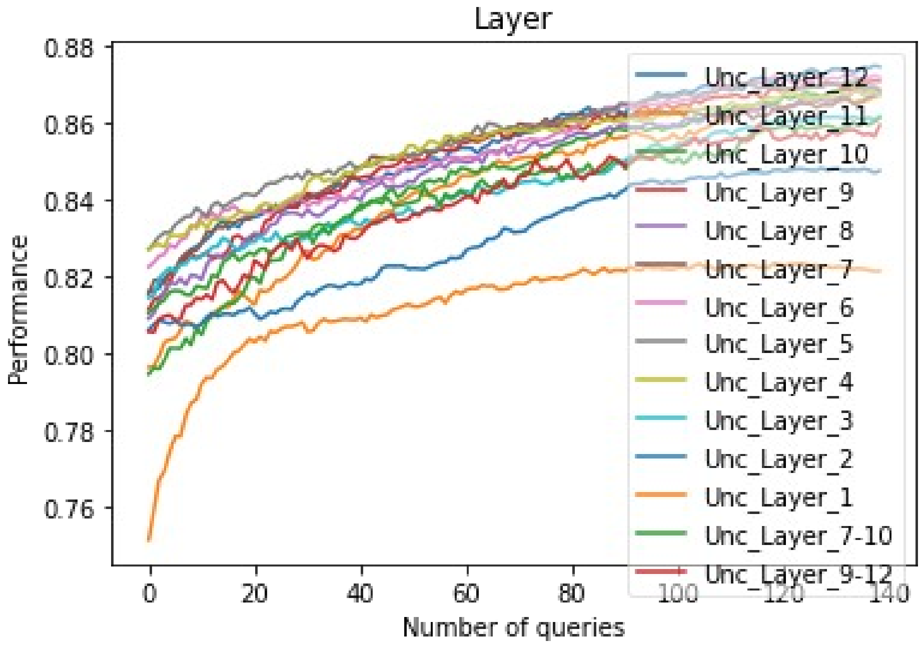


1	2	3	4	5	6	7	8	9	10	11	12	7.10	9.12
71.8	73.7	74.3	75.5	76.2	78.2	77.2	79.0	81.2	82.5	83.0	83.6	79.0	80.9

Results AG‘s News

Data-set

- 4 largest news categories
- Labels;
  World, Sport, Business, Science/Tech (balanced)

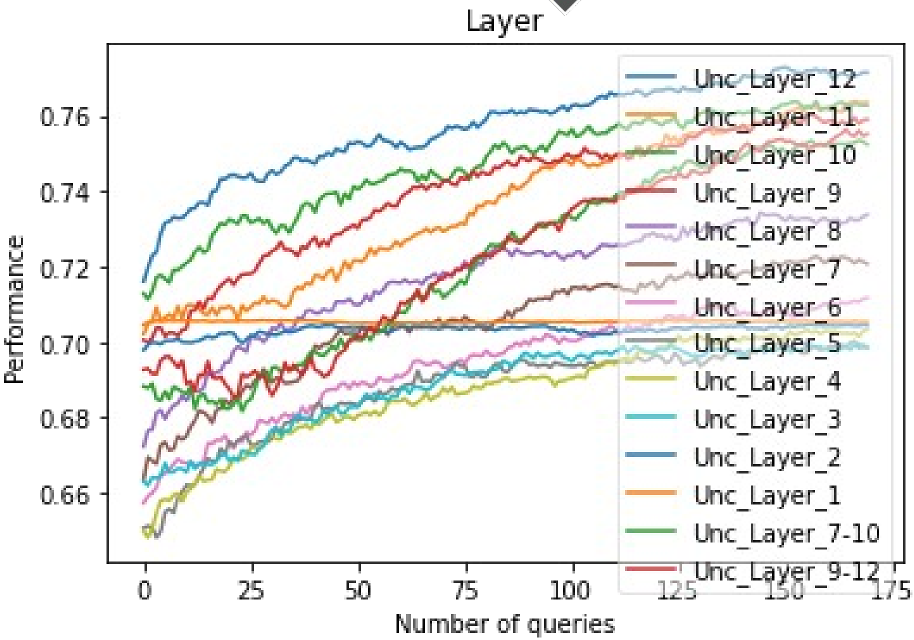


1	2	3	4	5	6	7	8	9	10	11	12	7.10	9.12
81.3	83.0	84.3	85.4	85.6	85.4	85.3	84.9	85.2	84.5	84.3	85.4	84.2	84.0

Results CoLa

Data-set

- Sentences from linguistic publications
- Labels; correct or not (imbalanced)



1	2	3	4	5	6	7	8	9	10	11	12	7.10	9.12
70.5	70.3	68.8	68.7	68.7	69.4	70.5	71.7	74.0	74.7	73.6	75.7	72.2	72.2

Conclusion

Conclusion

Dresden Database Systems Group

Preceding Work

- Focused on output layer
- Increased classification performance

Our work

- Active learning increased performance for all layers and combinations
- Low amount of labeled data
- Class imbalances
- Multinomial and binomial data
- Addition and concatenation
- Time constraints

Dresden Database Systems Group

Thanks! Questions?