Active Learning for Speech Emotion Recognition using Deep Neural Network

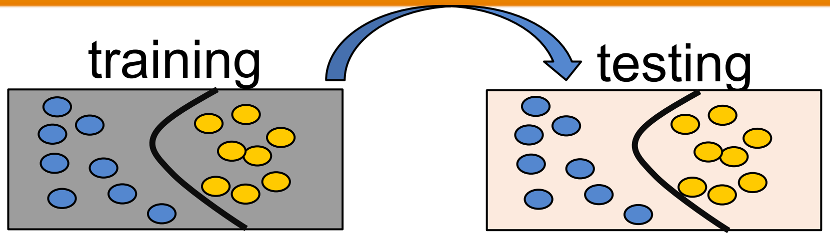
Mohammed Abdelwahab And Carlos Busso

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP Multimodal Signal Processing Laboratory

Generalization of Models

- **Mismatch between train and test conditions is one of the main barrier in speech emotion recognition**
- **Under ideal classification conditions**
  - The training and testing sets come from the same domain
- **Under real application conditions**
  - The training and testing sets come from the different domains
  - This leads to performance drop [Shami and Verhelst 2007, Parthasarathy and Busso 2017]



Training	Testing	Accuracy
Danish	Danish	64.90 %
Berlin	Berlin	80.70 %
Berlin	Danish	22.90 %
Danish	Berlin	52.60 %

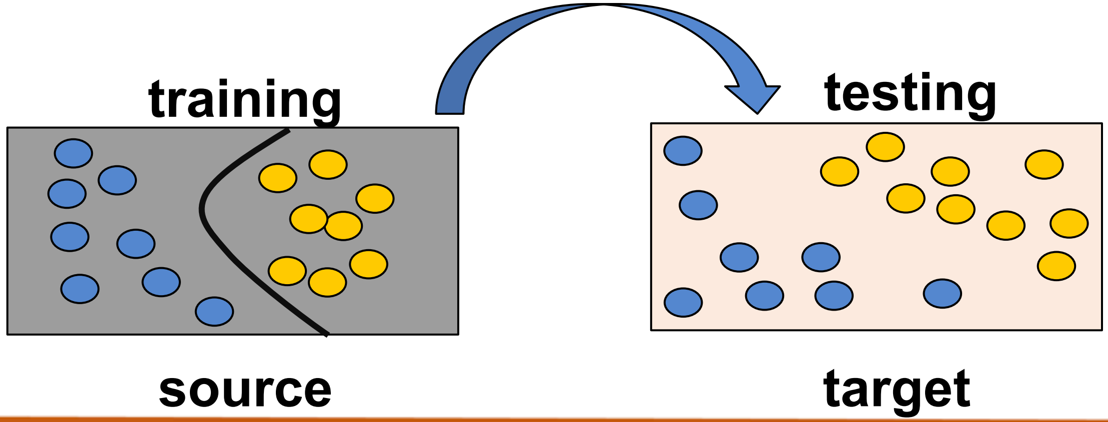
[Shami and Verhelst 2007]

Training	Arousal [CCC]	Valence [CCC]	Dominance [CCC]
In-corpus	0.764	0.289	0.713
Cross-corpus	0.464	0.184	0.451

[Parthasarathy and Busso 2017]

The Problem

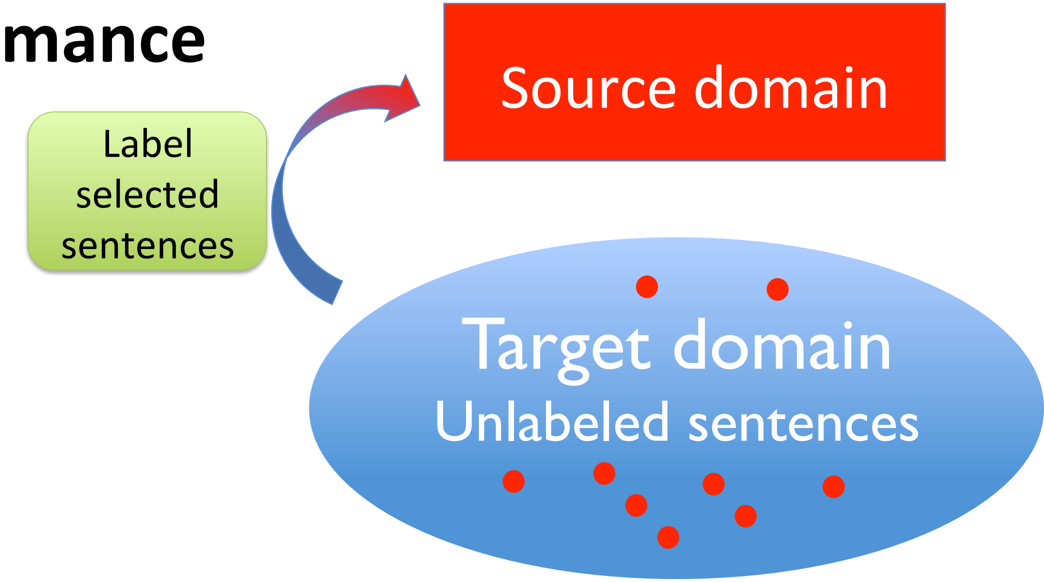
- **The performance of a classifier degrades if there is a mismatch between training and testing conditions**
  - Speaker variations, channels (environments, noise), language, and microphone settings
- **How to build a classifier that generalizes well?**
  - Minimize the discrepancy between the source and target domains



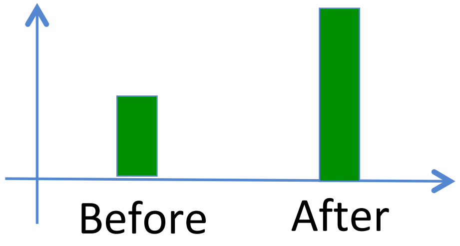
We explore this problem relying on **active learning**

Motivation

- **Active learning has been widely used to iteratively select training samples that maximizes the model's performance**
  - Not all the samples are equal



- **DNN pushes state of the art performance**
  - It requires vast amounts of labeled data
- **There is a need for scalable active learning approach for DNN**
  - Explore the approaches to identify most useful *N* samples

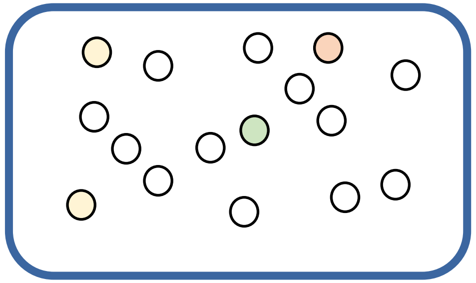


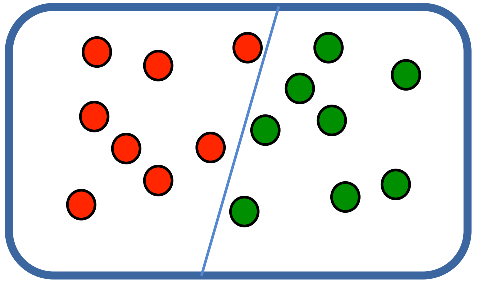
Related Work

- **Speech Emotion Recognition**
  - Use labelers' agreement to build uncertainty models [Zhang et. al. 2013]
  - Multi-view uncertainty sampling to minimize amount of labeled data [Zhang et. al. 2015]
  - Minimize annotations per sample using agreement threshold [Zhang et. al. 2015]
  - Minimize noise accumulation in self-training [Zhang et. al. 2016]
  - Adapt model with low confidence correctly classified samples [Abdelwahab & Busso 2017]
  - Combine Ensembles and Active learning to mitigate performance loss in new domain [Abdelwahab & Busso 2017]
  - Greedy sampling for Multi-task speech emotion linear regression [Wu & Huang 2018]
- **None of those approaches used Deep Neural networks**

Data Acquisition Functions

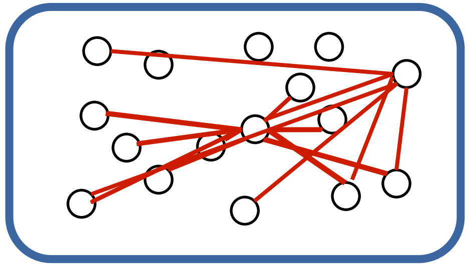
- There is no data acquisitions functions that work well in all scenarios
- Heuristic approaches where shown to work in practice
  - Greedy sampling
    - Label space
    - Feature space
    - Combination
  - Uncertainty sampling
    - Least confident samples
    - Margin
    - Entropy
    - Vote Entropy (Ensembles)
    - Dropout
  - Random sampling (baseline)

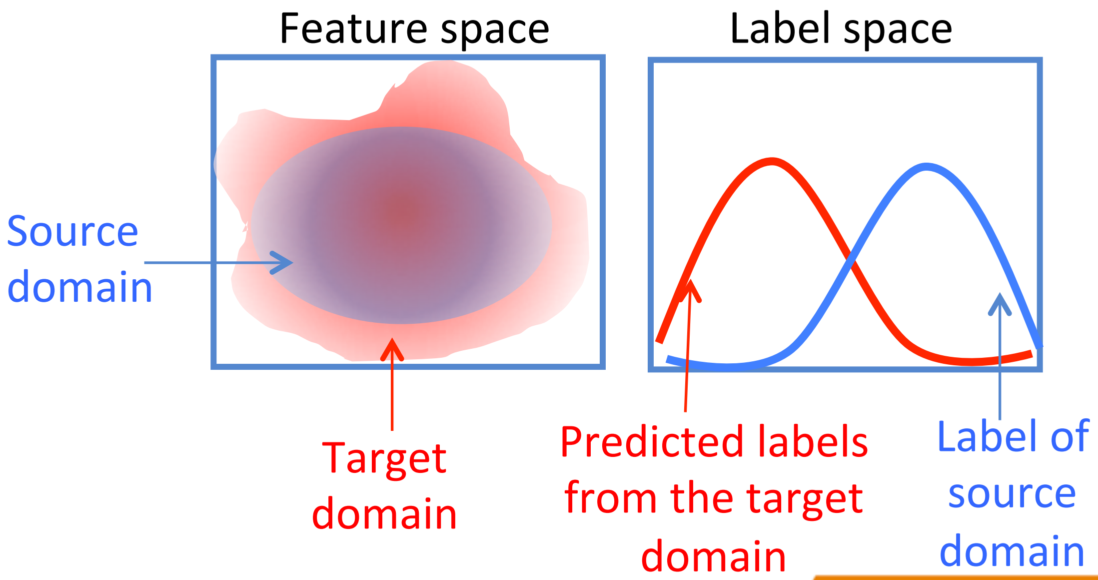




Greedy Sampling Approach

- **Greedy sampling for regression** **[Wu et al., 2019]**
  - maximize the diversity in the train set
  1. Select initial samples
     - Previously selected samples
  2. Compute distances
     - Features space $d_x^{i,j} = \|x_i - x_j\|_2$
     - Label space $d_y^{i,j} = |\hat{y}_i - y_j|$
     - Combination $d_{xy}^{i,j} = d_x^{i,j} d_y^{i,j}$
  3. Select *k* samples to annotate
  4. Update model and repeat

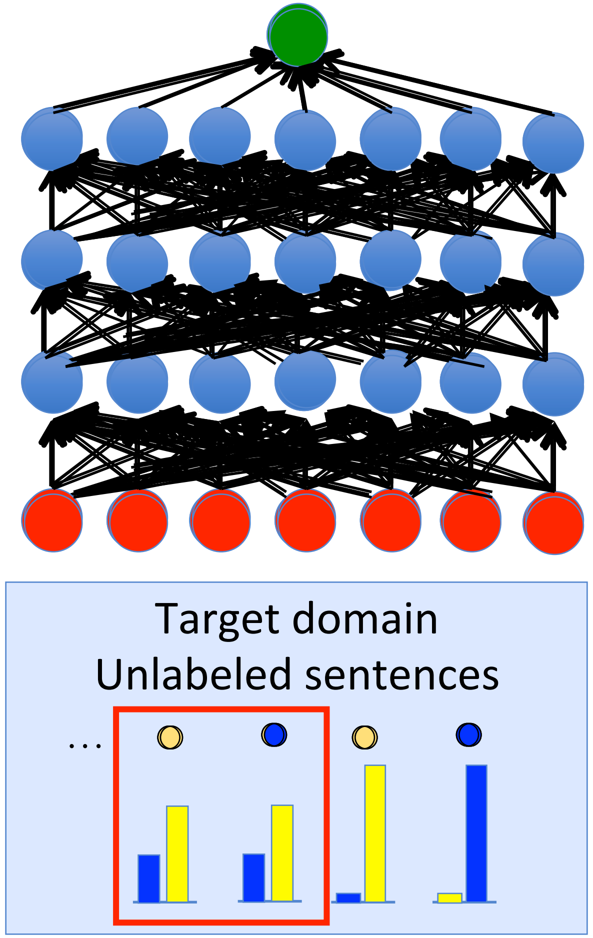




Uncertainty Sampling: Dropout

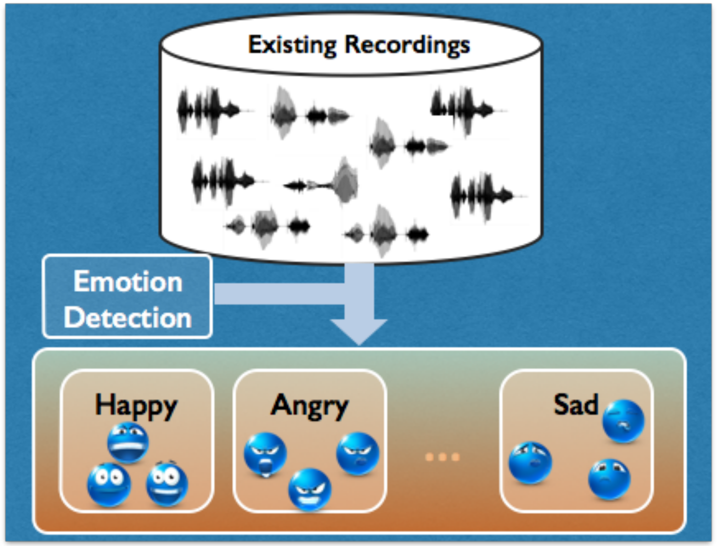
- **Dropout can approximate Bayesian inference [Gal et al., 2016]**
  - We can represent the models' uncertainty
  - Use different configurations of dropout, analyzing predictions per sample

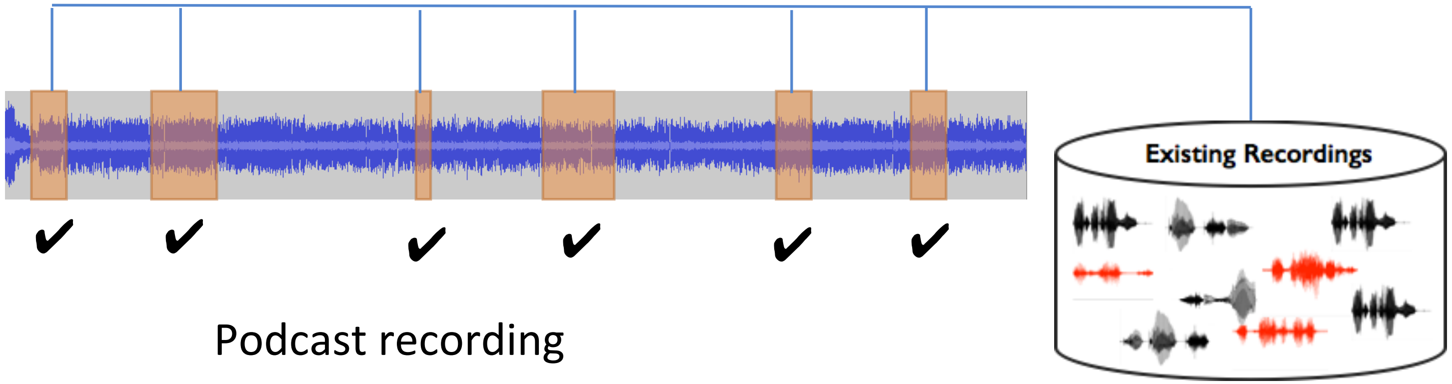
- **Goal: select samples that the existing model is the most uncertain across several dropout iterations**



The MSP-Podcast Database

- Use existing podcast recordings
- Divide into speaker turns
- Emotion retrieval to balance the emotional content
- Annotate using crowdsourcing framework

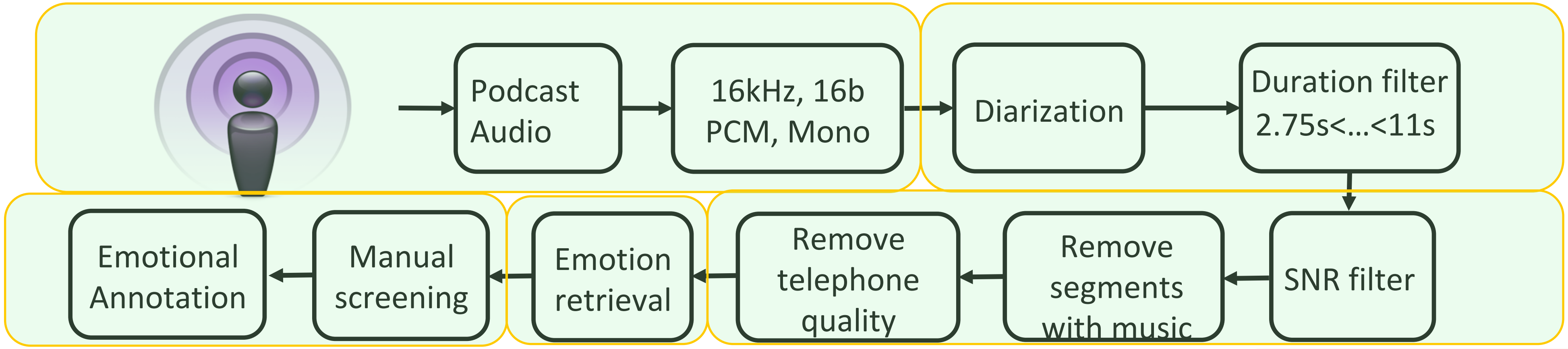




Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. To appear, 2018.

The MSP-Podcast Database

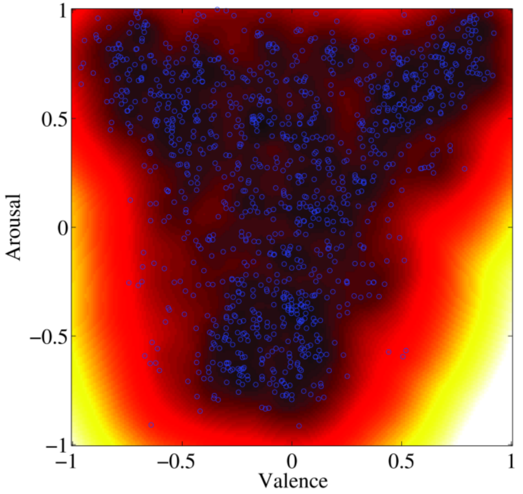
- **MSP-Podcast**
  - Collection of publicly available podcasts (naturalness and the diversity of emotions)
    - Interviews, talk shows, news, discussions, education, storytelling, comedy, science, technology, politics, etc.
  - Creative Commons copyright licenses
  - Single speaker segments, High SNR, no music, no phone quality
  - Developing and optimizing different machine learning framework using existing databases
    - Balance the emotional content
  - Emotional annotation using crowdsourcing platform



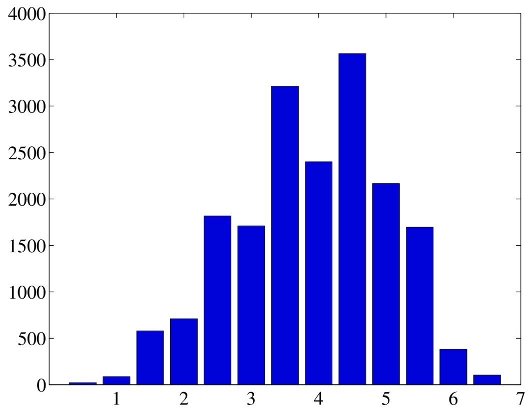
MSP-Podcast corpus version 1.1

With emotion labels:
22,630 sentences
(38h, 57m)

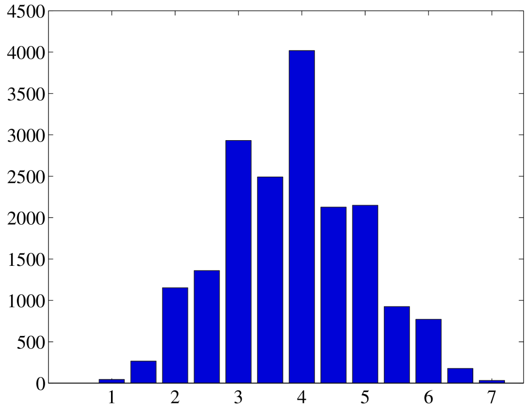
Segmented turns
152,975 sentences over 1,000 podcasts



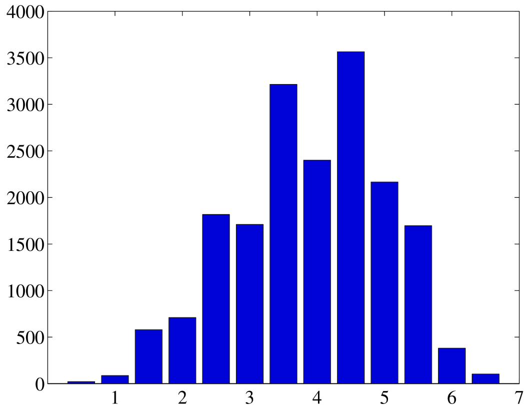
- Test set
  - 7,181 segments from 50 speakers (25 males, 25 females)
- Development set
  - 2,614 segments from 15 speakers (10 males, 5 females)
- Train set
  - remaining 12,830 segments



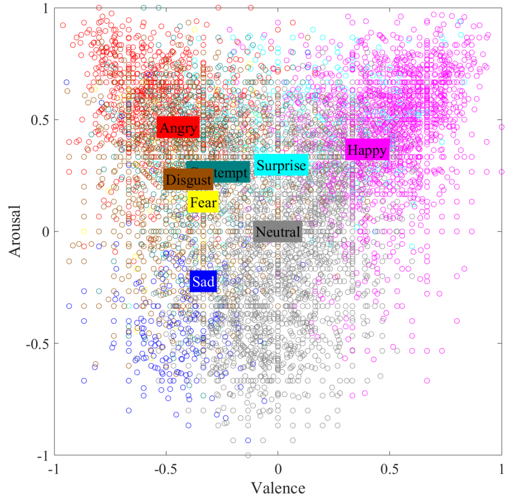
Arousal



Valence



Dominance



Acoustic Features

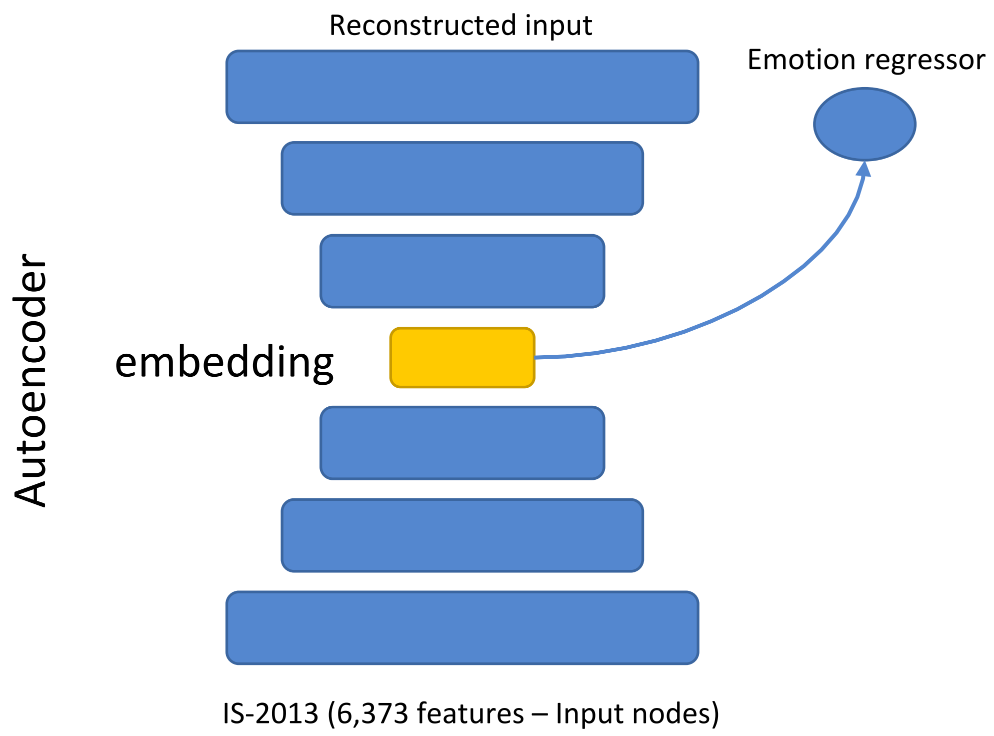
- **Interspeech 2013 Feature set**
  - 65 low level descriptors (LLD)
  - Functional are calculated on LLDs resulting in total of 6,373 features
  - Functionals include:
    - Quartile ranges
    - Arithmetic mean
    - Root quadratic mean
    - Moments
    - Mean/std. of rising/ falling slopes

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

Proposed Architecture

- **Multitask learning network:**
  - Primary task: emotion regression
    - concordance correlation coefficient (CCC)
  - Secondary task: feature reconstruction
    - Mean square error (MSE)

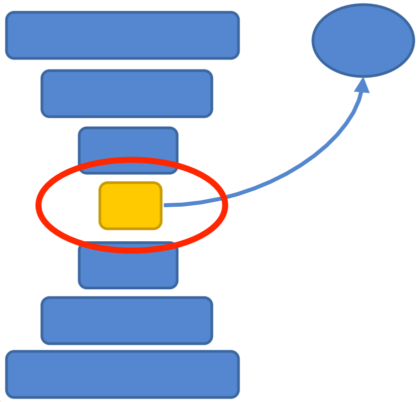
$$\mathcal{L} = \lambda_1 \underbrace{\frac{1}{N}\sum_{i=1}^{N} \|x - \hat{x}\|^2}_{\text{MSE}} + \lambda_2 \underbrace{\left[1 - \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}\right]}_{\text{CCC}}$$

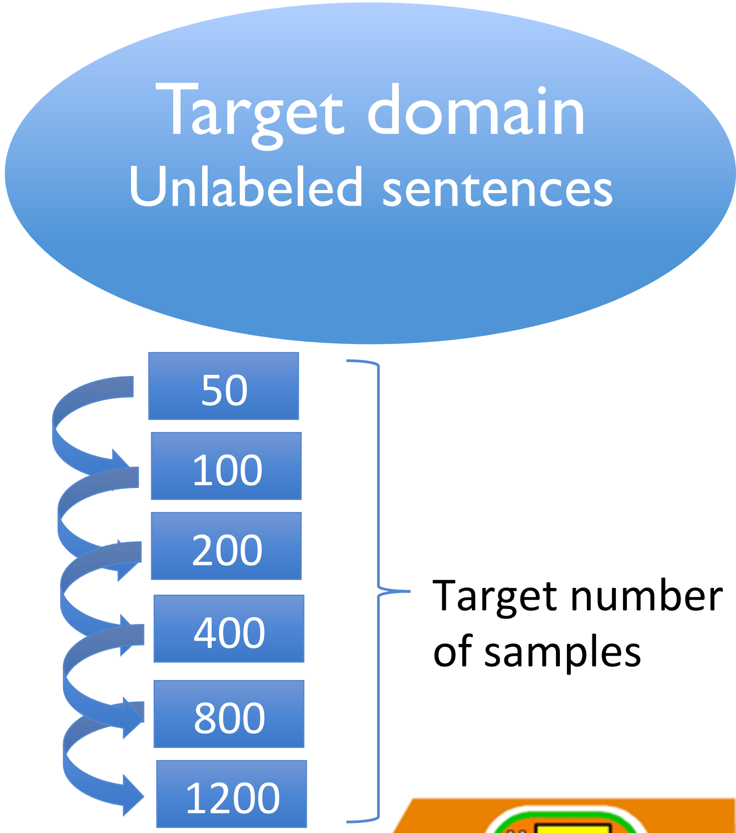


- Secondary task helps to generalize the model, especially with limited data

Experimental Settings

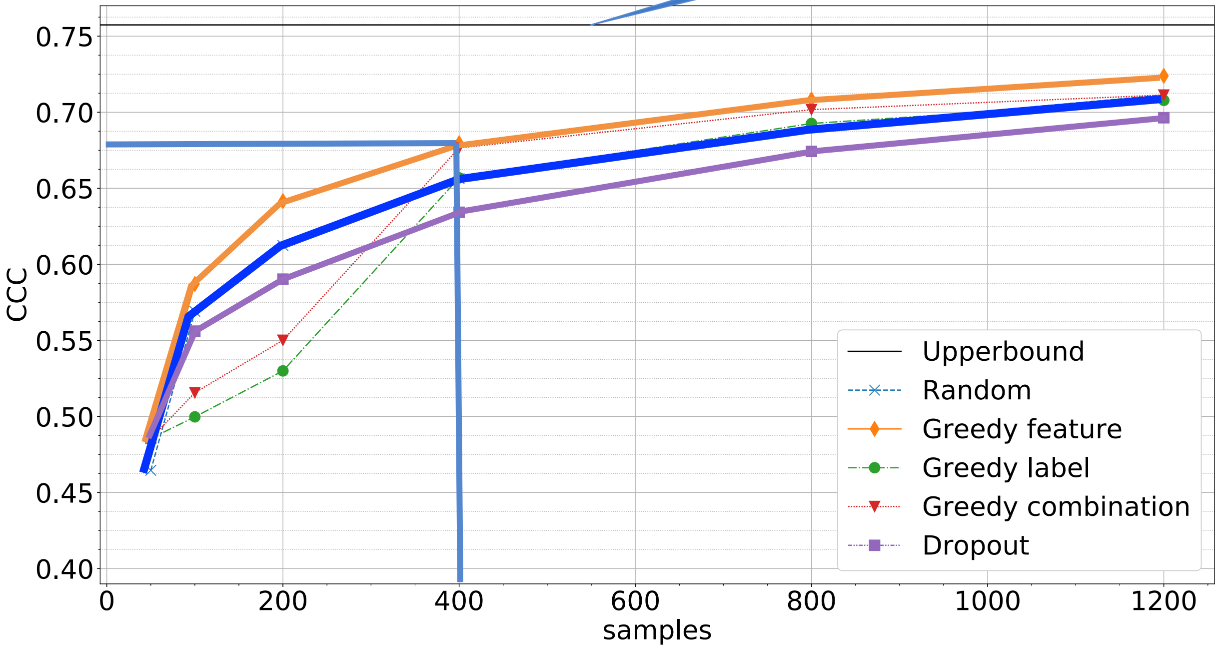
- **We consider 50, 100, 200, 400, 800, and 1200 samples**
- **Samples are selected based on the latest model**
- **We consider two starting points**
  - From scratch
  - Autoencoder trained on reconstruction loss only for 20 epochs
- **Results are the average of 20 trials**
- **Greedy sampling (feature space):**
  - Use embedding of the autoencoder to reduce the search space



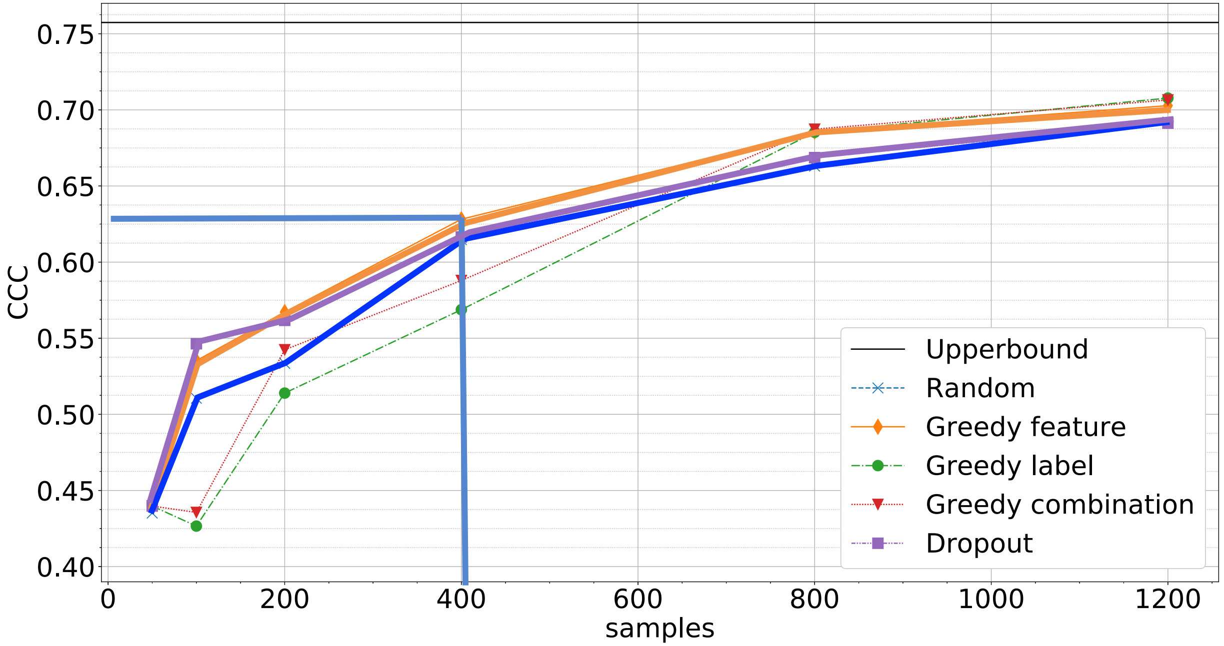


Results for Arousal

Within corpus performance



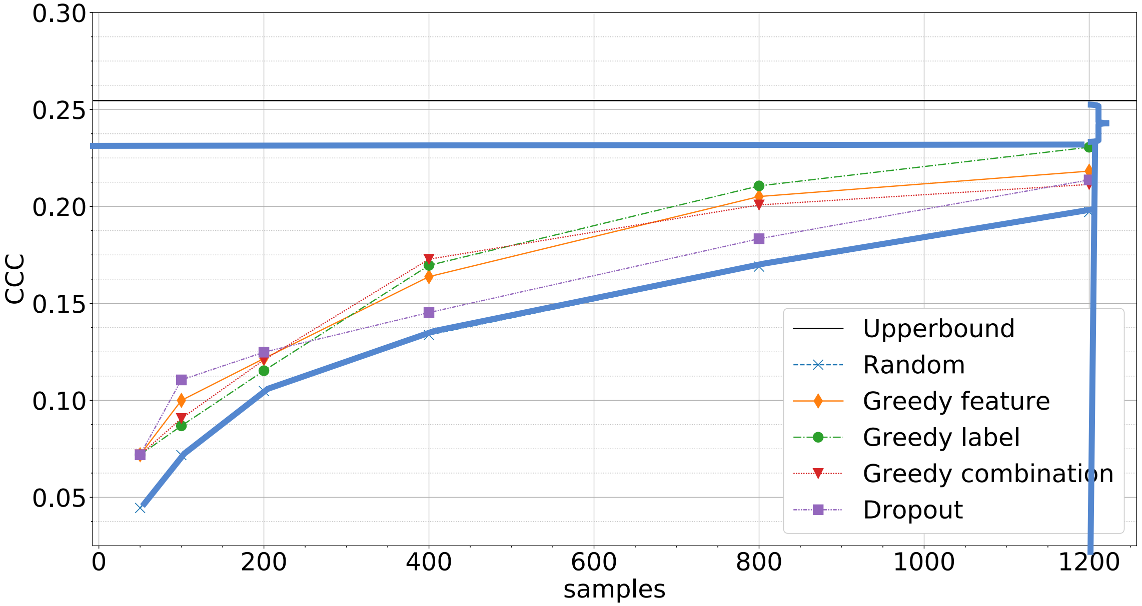
Pretrained encoder



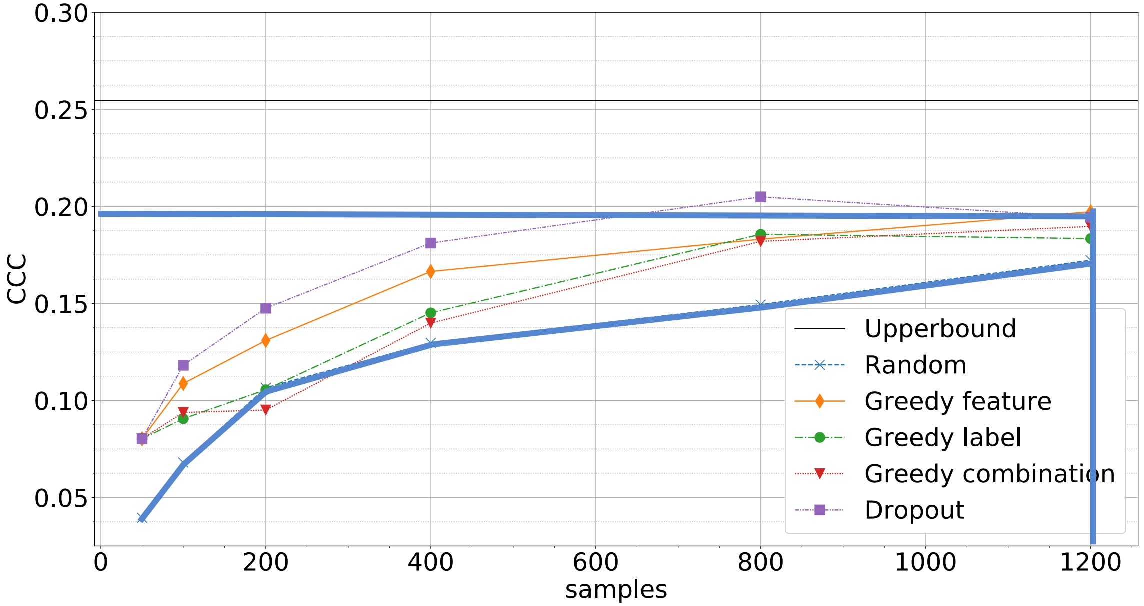
From scratch

- **Observation**
  - Greedy feature leads to better performance
  - Dropout is not as effective
  - Random approach best methods as we add more data
  - Pretrained encoder helps with limited samples

Results for Valence



Pretrained encoder



From scratch

- **Observations**
  - Pretrained autoencoder helps to achieve better performance
  - We approach within corpus performance with only 10% of the training data
  - Random sampling is less effective

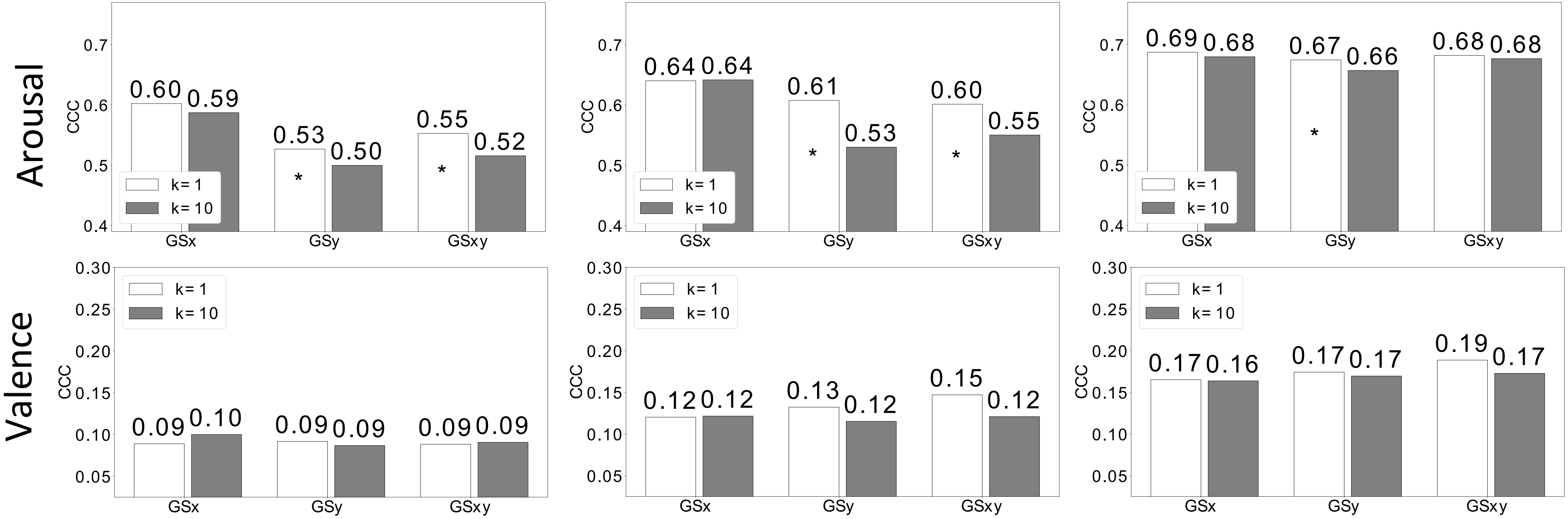
Statistical Significance

	Arousal				Valence			
# samples	100	200	400	800	100	200	400	800
Random Sampling	0.57	0.61	0.66	0.69	0.07	0.10	0.13	0.17
Greedy Feature	0.58	**0.64**	**0.68**	**0.71**	**0.10**	0.12	**0.16**	**0.21**
Greedy Label	0.50	0.53	0.66	0.69	0.09	0.12	**0.17**	**0.21**
Greedy Combination	0.52	0.55	0.68	0.70	0.09	0.12	**0.17**	**0.20**
Dropout	0.56	0.59	0.63	0.67	**0.11**	0.12	0.15	0.18

Bold: statistically significant improvements over random sampling

- **Observations**
  - Greedy sampling in feature space almost always better than random sampling
  - Dropout was not as effective

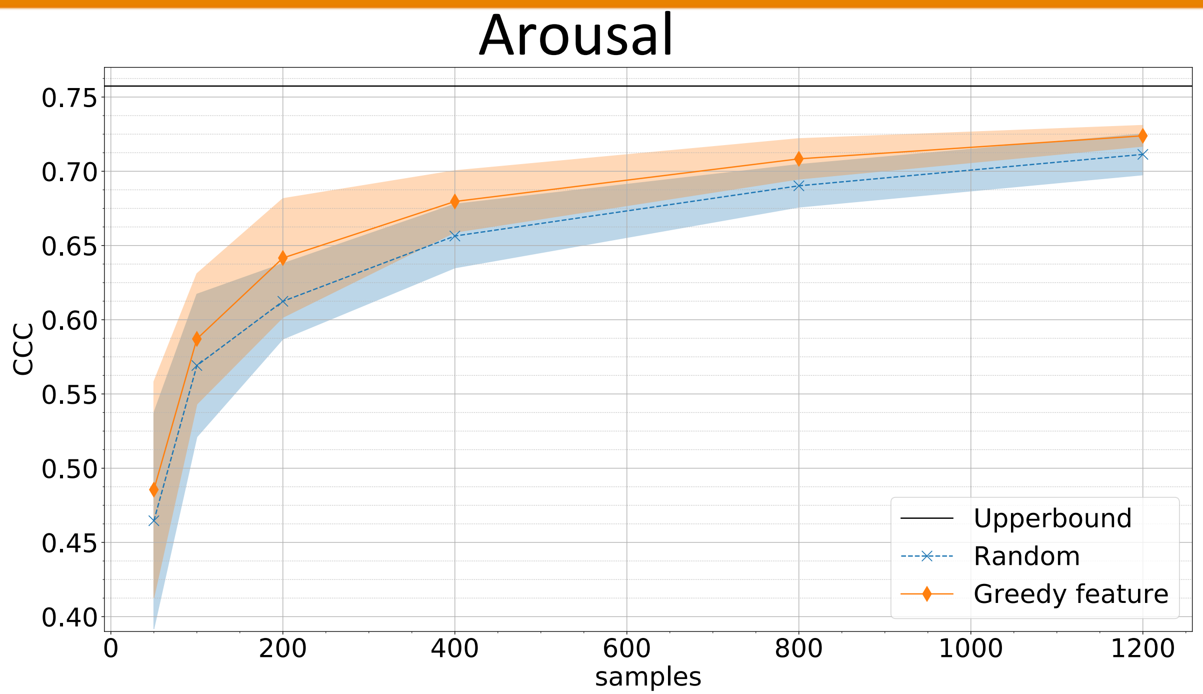
Sensitivity to *k* (how often we update the model)

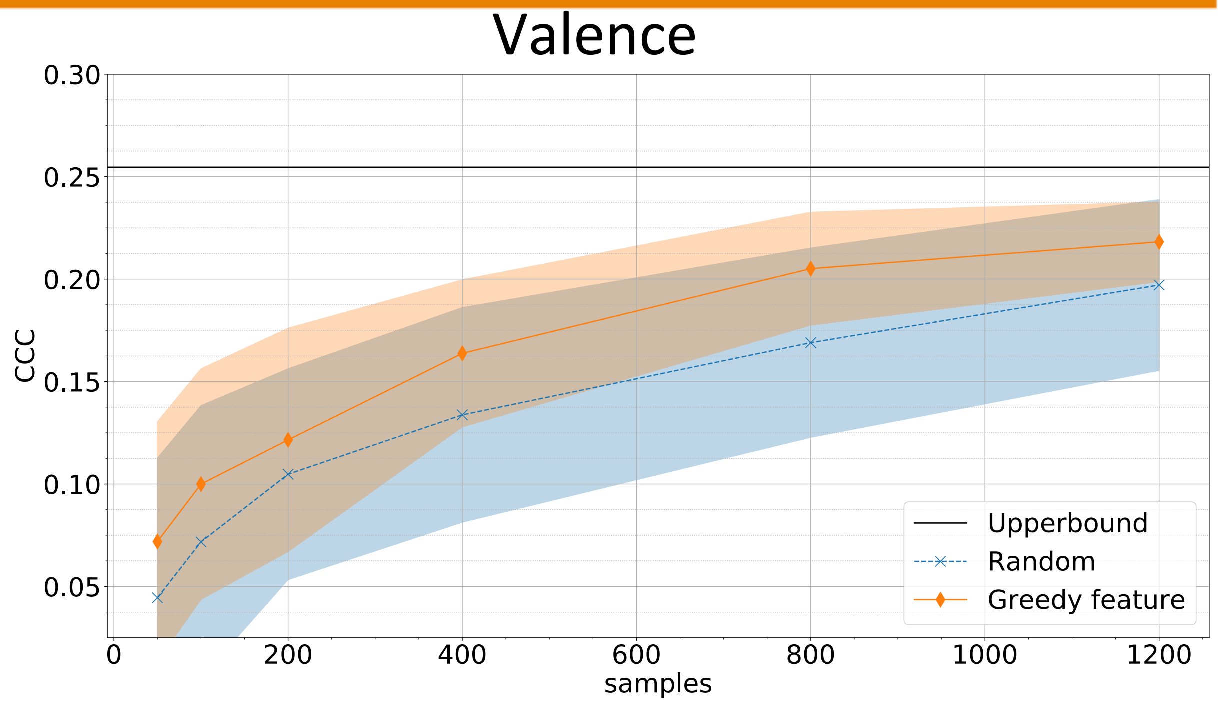


100 samples | 200 samples | 400 samples

- **Multitask autoencoder framework with greedy methods**
  - No statistical difference with *k* = 1 and *k* = 10
  - Method is not sensitive to this parameter (reduce complexity)

Consistency of the Results





- **20 results starting with different initializations**
  - Greedy sampling on the feature space versus random sampling
  - Standard deviation of the CCC values achieved by the greedy sampling method decreases faster as the sampling size increases
    - More consistent than random sampling

Conclusions

- **Greedy sampling achieves higher performance with lower variance compared to random sampling**
- **Greedy sampling in label space depends on model's performance**
- **As we introduced more data, the differences in performance across data acquisition functions reduce**
- **Reduce computation cost:**
  - Calculate the distance in embedding with lower dimensions
  - Set adequate value of *k* reduces the frequency of model updates

- **Future Work**
  - Combine active learning with curriculum learning
  - Consider new acquisition functions that scale well

Thank you

- **This work was funded by NSF CAREER award IIS-1453781**

Interested on our research?
msp.utdallas.edu