Ranking Emotional Attributes With Deep Neural Networks

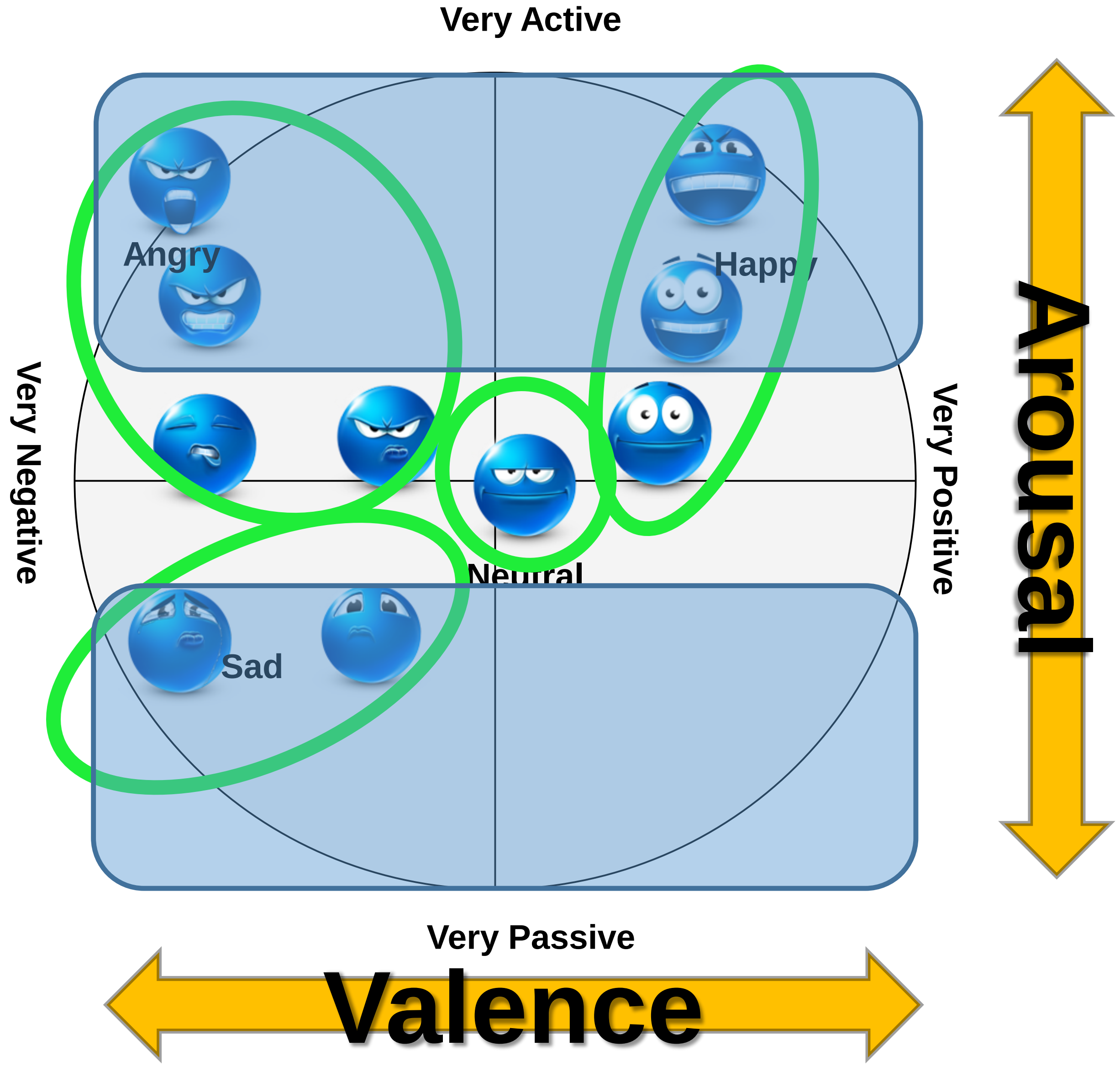
Srinivas Parthasarathy, Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

March 8, 2017

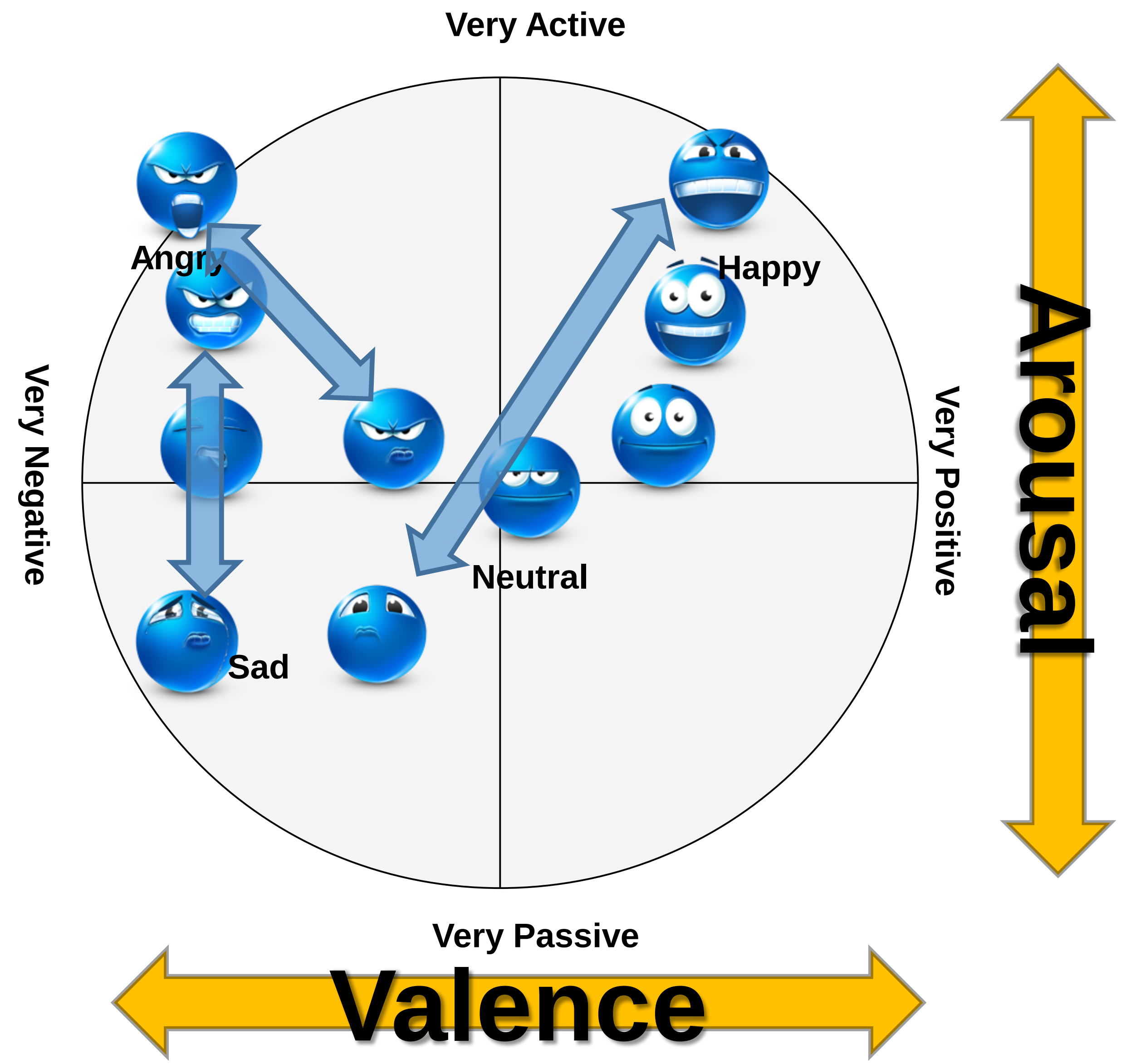
Motivation

- Emotion recognition systems can be trained to
  - Classify discrete categories such as Happy, Neutral, Angry etc.
  - Classify or predict values of emotional attributes such as
    - Arousal (passive vs active)
    - Valence (positive vs negative)



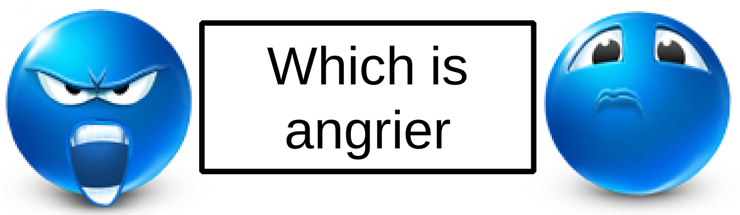
Motivation

- Humans are better at relative comparisons than absolute values
- Rank emotional attributes rather than absolute classification/regression
- Appealing for Emotional Retrieval tasks
  - Rank order aggressive behavior
  - Retrieve target behaviors with given emotions

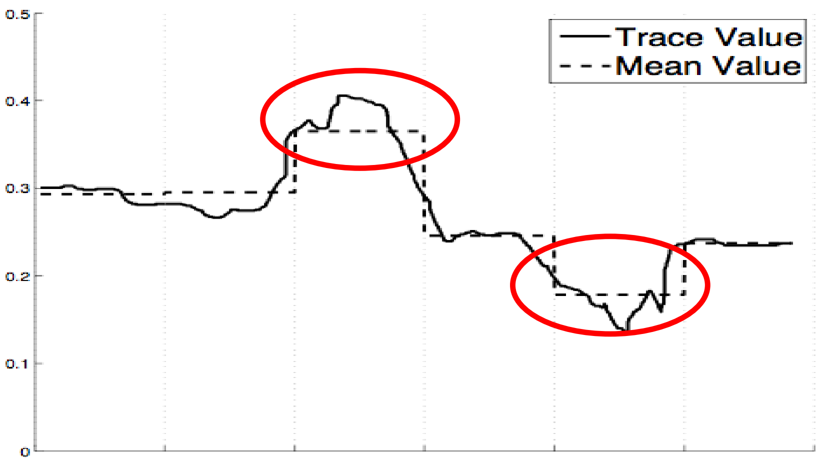


Related Work

- Commonly formulated as comparisons between pairs of samples
- Rankers for categorical emotions (e.g. angry rankers) [Cao et al. 2012, 2014]
    - Pairs formed between preferred emotion and other emotion



- Preference learning methods were used to learn from continuous ratings [Martinez et al. 2014]
- Alternative framework to study trends where raters agreed [Parthasarathy et al. 2016]

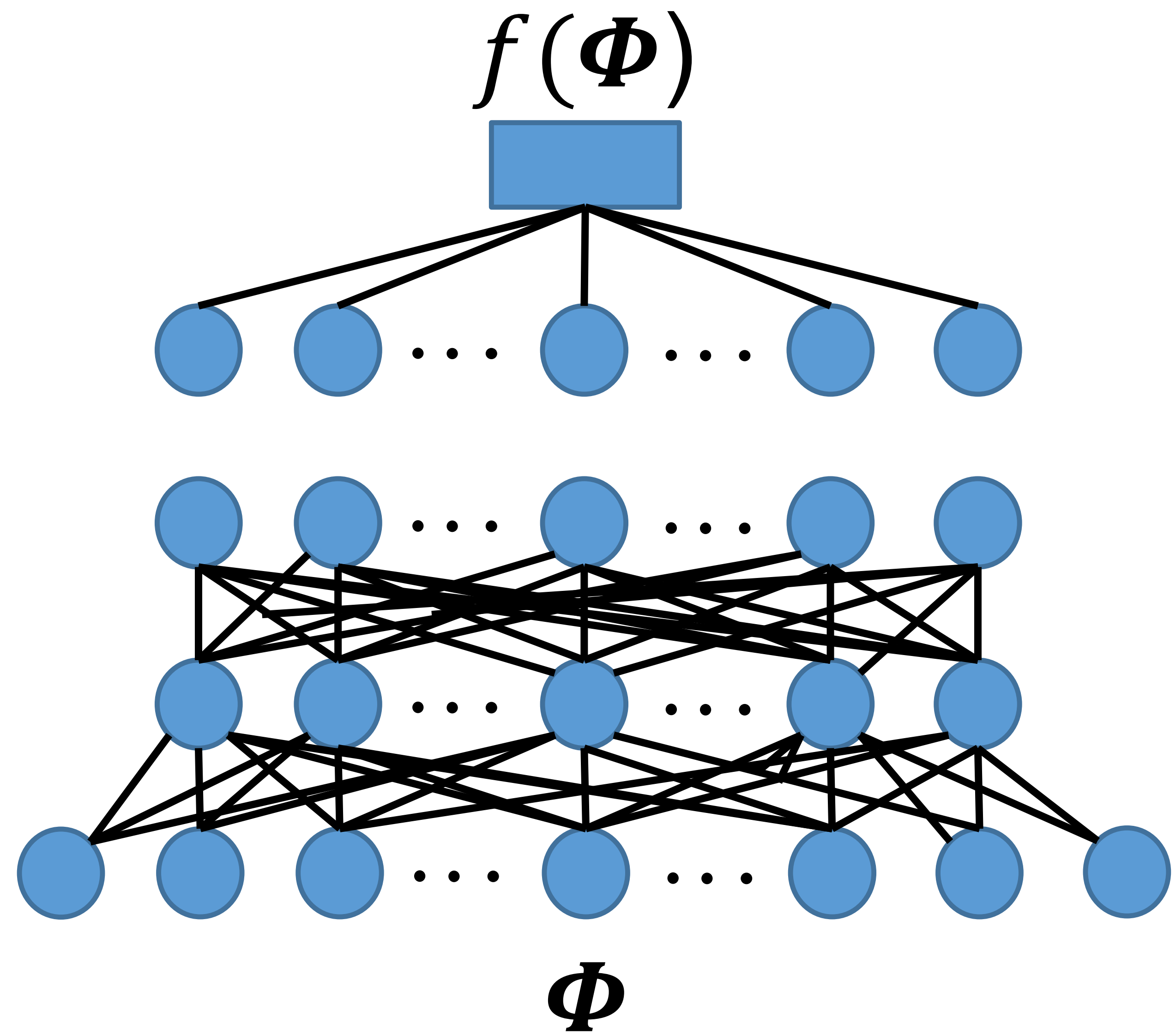


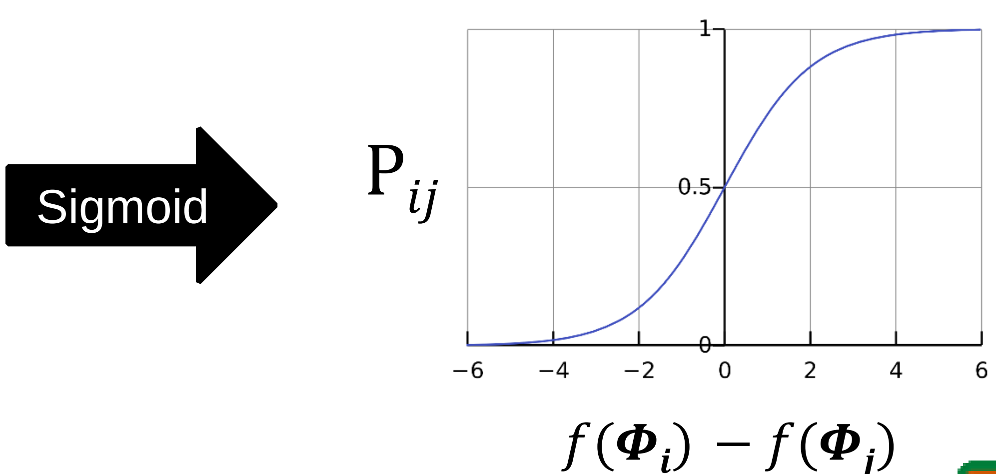
Contributions

- We rank order emotional attribute
- None of the previous studies have focused on using neural net learning techniques for preference learning
- We utilize a neural network framework for preference learning – RankNet
- To our knowledge, this is the first study that uses neural networks for ranking emotional attributes

RankNet

- Given: samples i, j, with features $\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j$
- Goal: Find f that learns the probability, P_{ij}, that $i \gg j$
- Neural network learns the function f, which maps feature vector $\boldsymbol{\Phi}$, to $f(\boldsymbol{\Phi})$
- Probabilistic framework
  - $\mathrm{P}_{ij} \equiv \dfrac{1}{1+e^{-\sigma(f(\boldsymbol{\Phi}_i)-f(\boldsymbol{\Phi}_j))}}$





RankNet

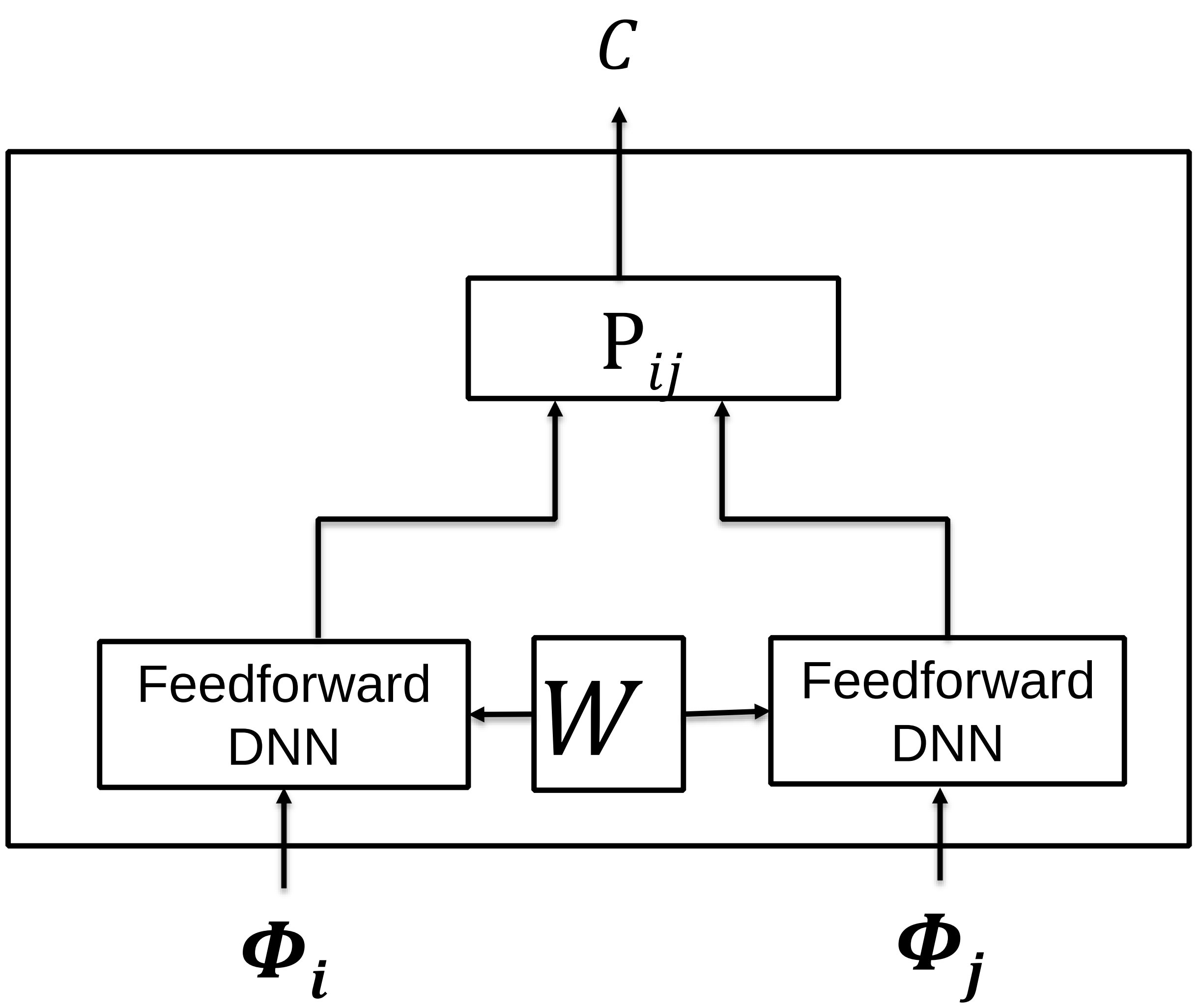
- Ideal probabilities $\overline{\mathrm{P}_{ij}}$ is set according to the preference in pairs of samples.
  - $\overline{\mathrm{P}_{ij}}$ = 0 if $j \gg i$
  - $\overline{\mathrm{P}_{ij}}$ = 1 if $i \gg j$
- Cross entropy is then used as the cost function to measure deviation of model

$$\boxed{C = -\overline{\mathrm{P}_{ij}} log(\mathrm{P}_{ij}) - (1 - \overline{\mathrm{P}_{ij}}) log(1 - \mathrm{P}_{ij})}$$

- Simplifies to
  - $C = log\left(1 + e^{-\sigma(f(\boldsymbol{\Phi}_i) - f(\boldsymbol{\Phi}_j))}\right)$ when $\overline{\mathrm{P}_{ij}}$ = 1
  - $C = log\left(1 + e^{-\sigma(f(\boldsymbol{\Phi}_j) - f(\boldsymbol{\Phi}_i))}\right)$ when $\overline{\mathrm{P}_{ij}}$ = 0

RankNet Framework

- The neural network for RankNet can be modeled with a **Siamese** architecture
- Features of pairs of samples are fed at the input
- Train two identical neural networks that share all parameters



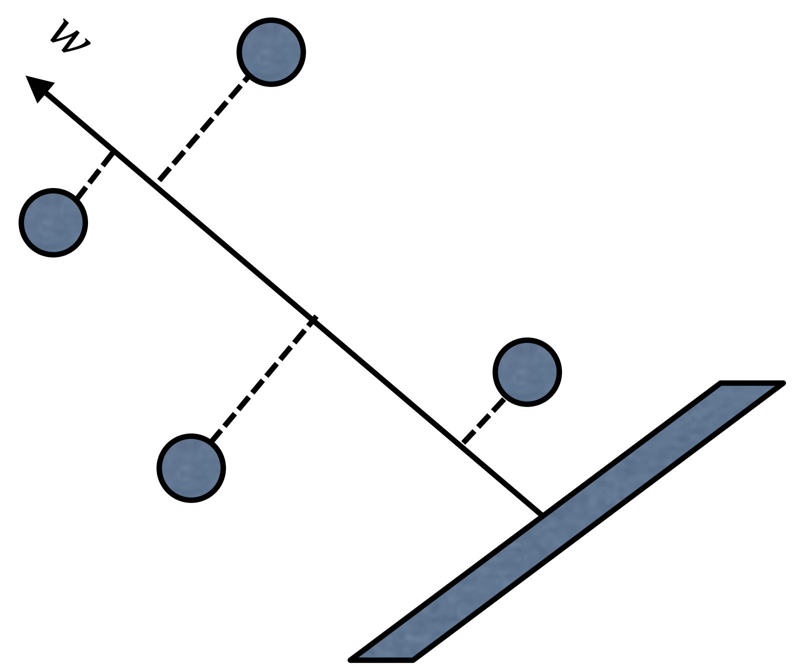
Baselines

- RankSVM framework for recognizing emotional attributes [Lotfian & Busso 2016]

- Given: $i \gg j$ goal is to

$$\min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i,j} \xi_{i,j}$$

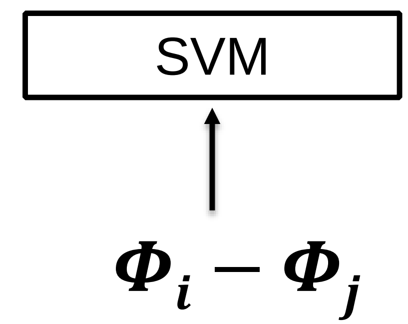
$$s.t\ \langle w, (\boldsymbol{\Phi}_i - \boldsymbol{\Phi}_j) \rangle \geq 1 - \xi_{i,j}\ and\ \xi_{i,j} \geq 0$$



- Reduced to binary classification with $\boldsymbol{\Phi}_i - \boldsymbol{\Phi}_j$

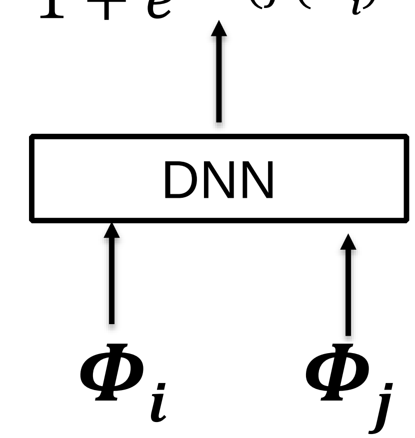
Differences

- RankSVM
  - Input is restricted to difference between features $\boldsymbol{\Phi}_i - \boldsymbol{\Phi}_j$
  - Large margin classifier
    - Redundant data can be removed
    - Performance does not increase with data [Lotfian & Busso 2016]
  - Kernel methods for non-linear classification



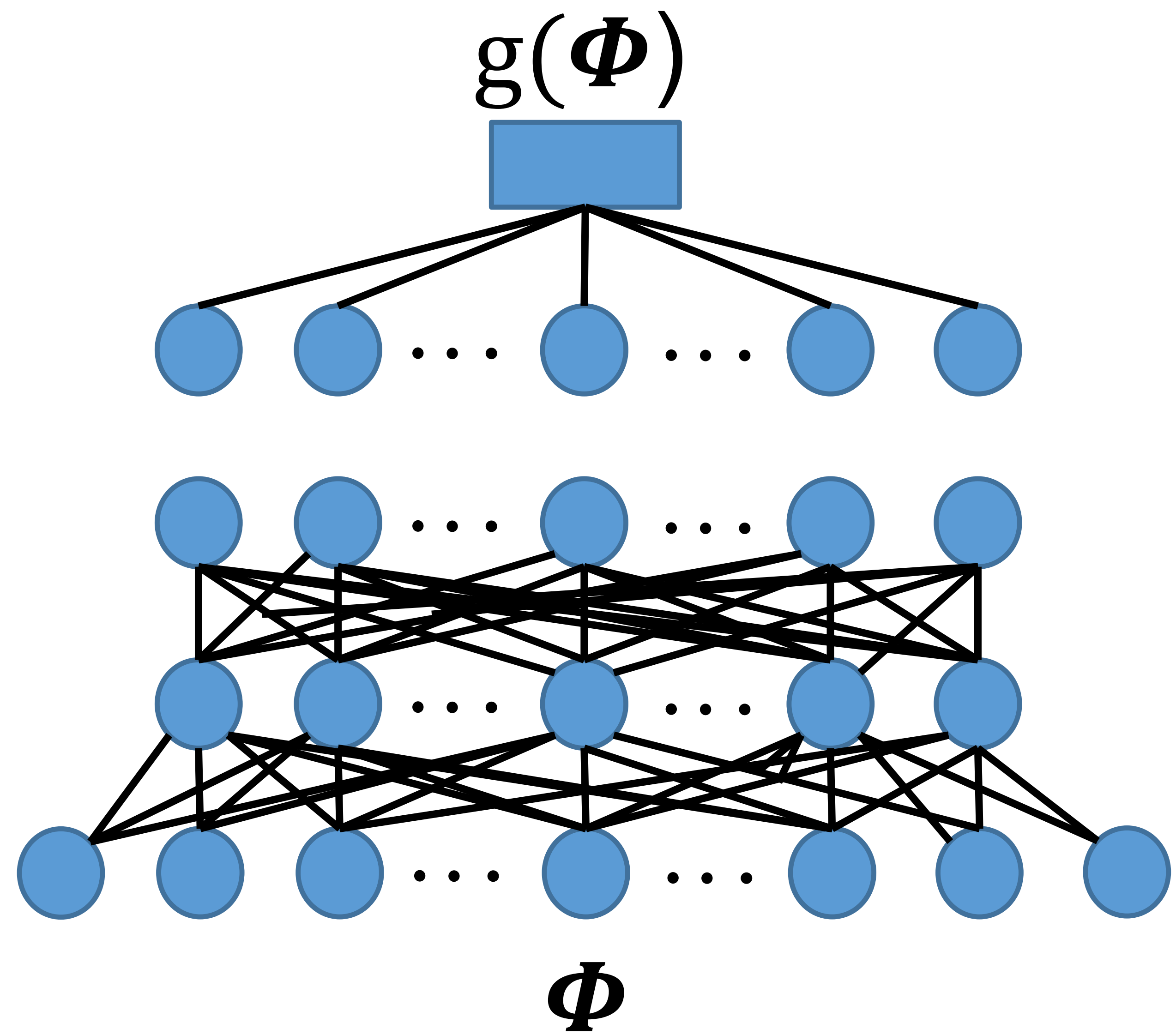
- RankNet
  - Features $\boldsymbol{\Phi}$ individually fed with no restrictions
  - Learns a non-linear mapping $f(\boldsymbol{\Phi})$
    - Optimized for pairs of samples
  - Highly data and parameter dependent

$$\mathrm{P}_{ij} \equiv \frac{1}{1 + e^{-\sigma(f(\boldsymbol{\Phi}_i) - f(\boldsymbol{\Phi}_j))}}$$



Baselines

- DNNRegression: Regression using DNNs
- No relative comparisons
- Use scores, $g(\boldsymbol{\Phi})$ to rank order sentences



Databases

- <u>Train:</u> USC-IEMOCAP
  - 12 hours of conversational recordings from 10 actors in dyadic sessions
  - Sessions consists of emotional scripts as well as improvised interactions
  - All speaking turns annotated for emotional attributes by two raters on a scale of 1-5
  - Arousal, Valence and Dominance
- <u>Test:</u> MSP-IMPROV
  - Improvisation between actors (12 actors)
  - Contains 8,438 speaking turns
  - Annotated by novel crowdsourcing methods on a scale of 1-5 by at least 5 raters
  - Arousal, Valence and Dominance

IEMOCAP

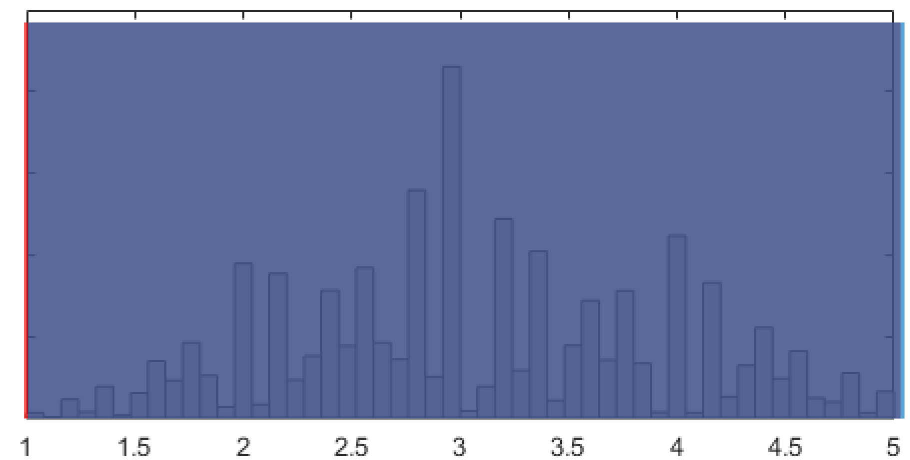
MSP-IMPROV

Experimental Settings

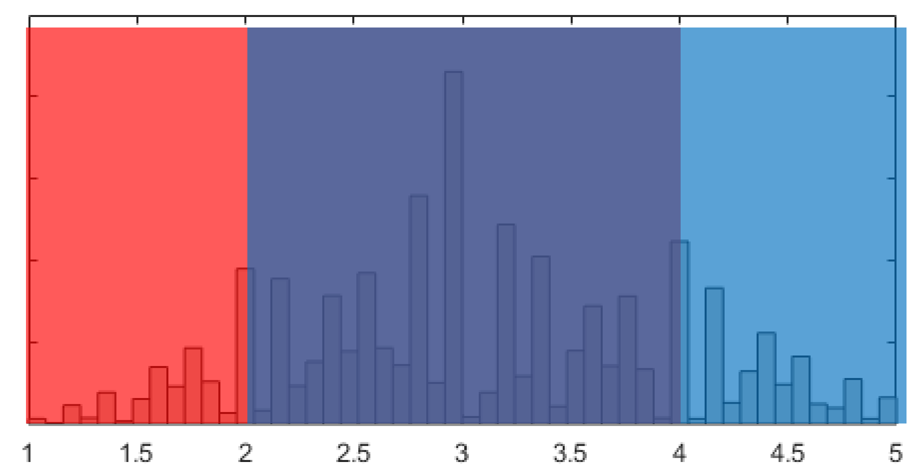
- Acoustic Features
  - Geneva Minimalistic Acoustic Parameter Set [Eyben et al. 2016]
    - Minimalistic features selected based on their performance in previous studies
    - Extended set – 88 features
    - Reproducibility (no feature selection)
    - Theoretical significance
- All DNN architectures include
  - 2 hidden layer, feed forward architecture 256 nodes each
  - Sigmoidal activation function
  - Stochastic Gradient Descent, learning rate of 10^{-4} for 100 epochs

Experimental Settings

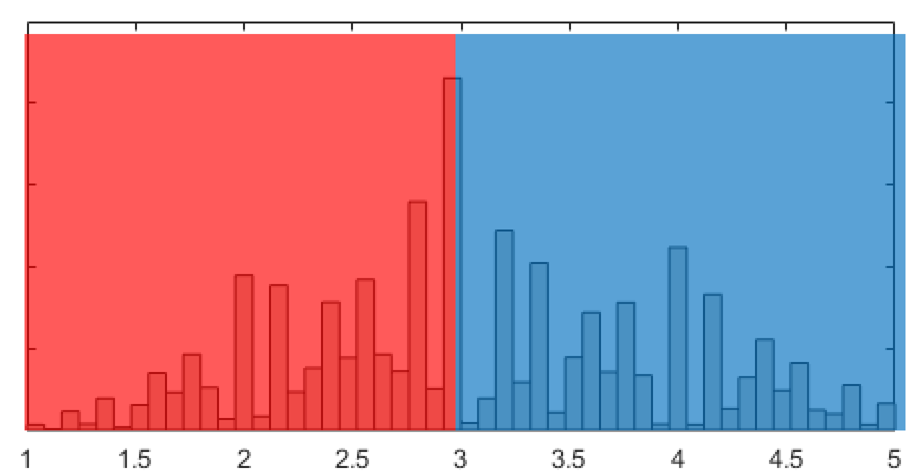
- Relative labels: consider samples separated by margin t
- $|S1_{arousal} - S2_{arousal}| > t$
- Tradeoff between t and data size
  - $\overset{\text{increase}}{\Uparrow} t \Leftrightarrow \overset{\text{increase}}{\Uparrow}$ reliability $\Leftrightarrow \overset{\text{decrease}}{\Downarrow}$ data
- RankSVM: $t = 1.0$ for arousal and dominance $t = 0.9$ for valence[Lotfian & Busso 2016]
- For RankNet we study the performance for $t \in \{0,1,2,3\}$
- Regression has no relative scores



t = 0



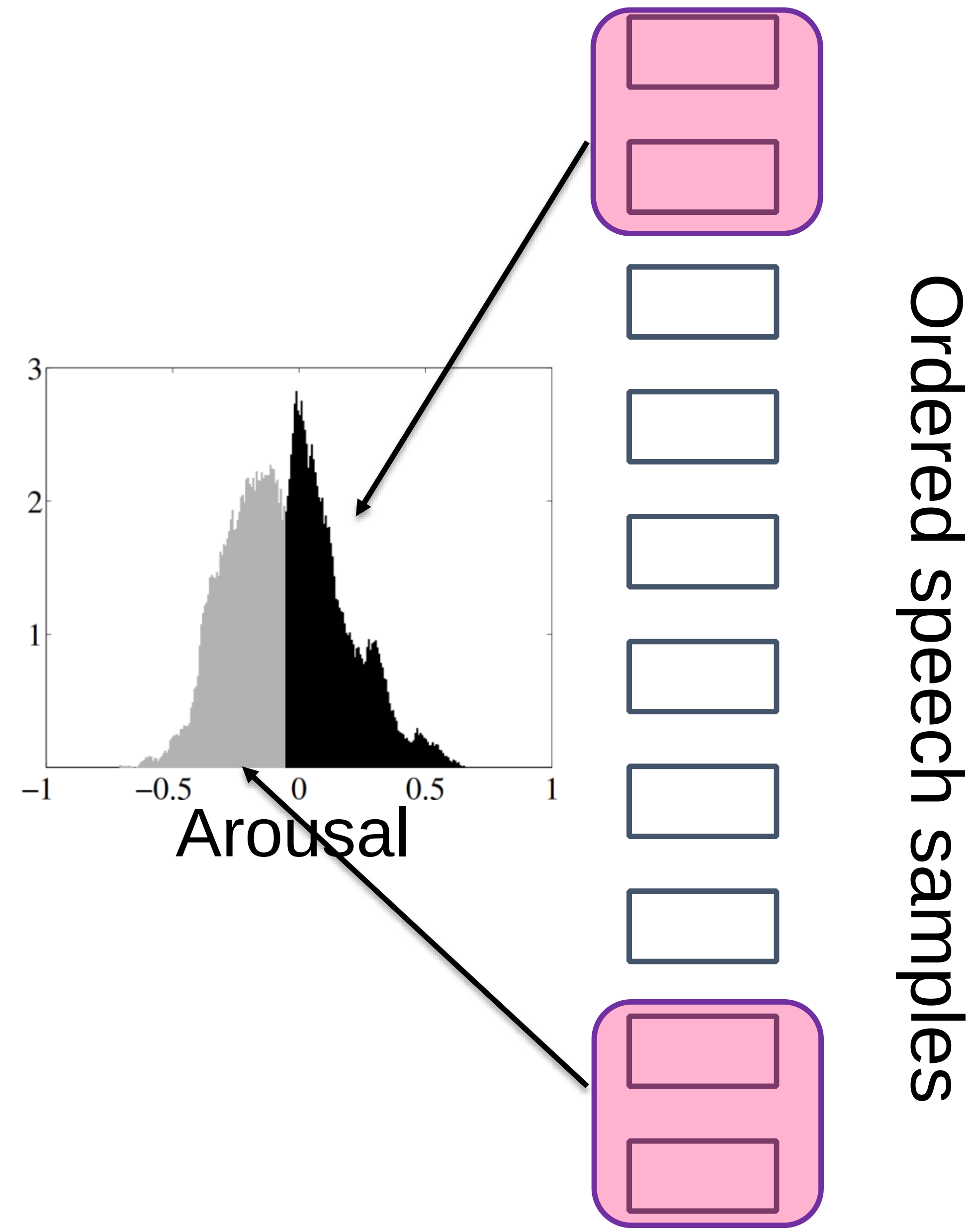
t = 1



t = 2

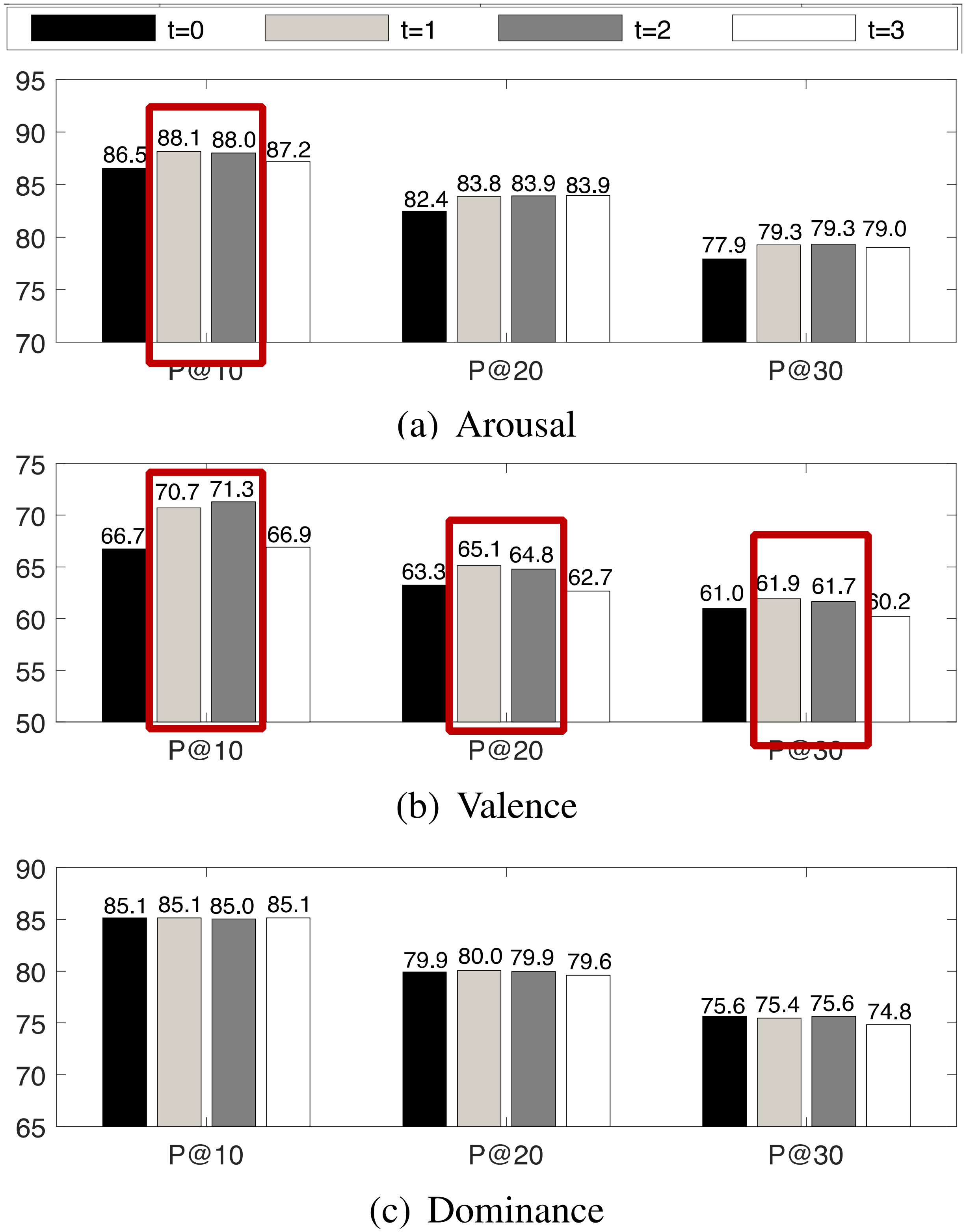
Evaluation

- Precision at k ($P@k$)
  - Measures the precision at retrieving k % of the samples from top and bottom
  - Ground truth is split into high and low classes about the median
  - Evaluate success in retrieving samples on the correct side of the split



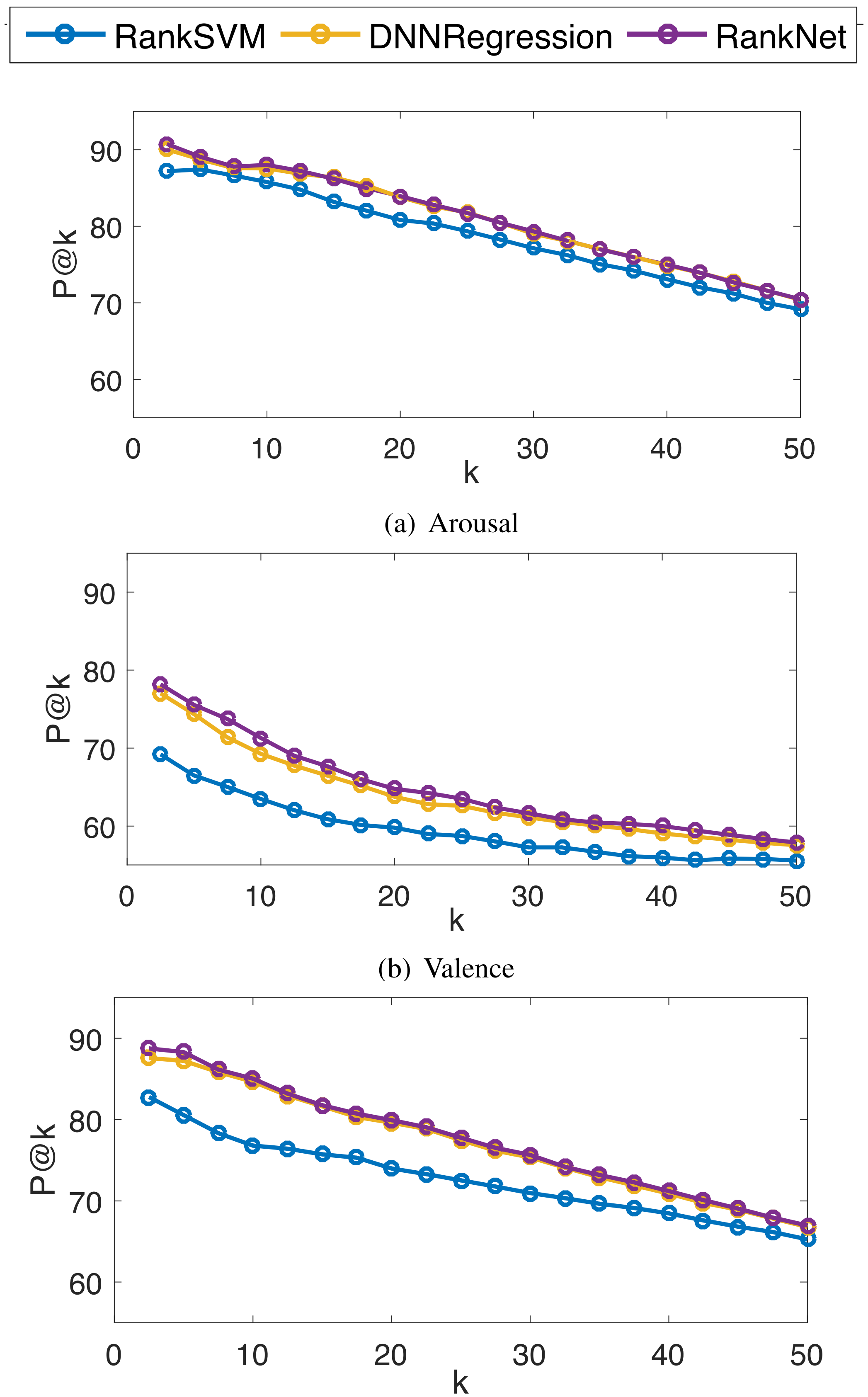
Effect of Margin on RankNet

- Attributes annotated on scale of 1-5
- P@10, P@20, P@30
- We see improvement for $t = 1{,}2$ but decrease $t = 3$.
- Use $t = 2$ for RankNet



Comparisons

	RankSVM	**RankNet**	DNNRegression
Arosual			
P@10	85.77	**88.02**	87.54
P@20	80.81	**83.93***	83.72*
P@30	77.15	**79.32***	79.02*
Valence			
P@10	63.46	**71.29***	69.28*
P@20	59.79	**64.77***	63.76*
P@30	57.26	**61.66***	61.13*
Dominance			
P@10	76.79	**86.15***	84.67*
P@20	73.97	**79.94***	79.61*
P@30	70.95	**75.65***	75.33*



(a) Arousal

(b) Valence

(c) Dominance

* Denotes Statistical Significance over RankSVM (population proportion)

Results

- Kendall's Tau Coefficient τ
  - Correlation between the two ordered lists [-1,1]

	RankSVM	**RankNet**	DNNRegression
Arousal	0.36	**0.41***	0.41*
Valence	0.08	**0.14***	0.13*
Dominance	0.28	**0.35***	0.34*

- RankNet and DNNRegression outperform RankSVM in all cases for $P@k$ and Kendall's τ
- Kendall's τ values are better than those reported in previous studies
  - τ values $\approx$ 0.02 for Arousal, 0.05 valence$_{\text{[Martinez et al. 2014]}}$

Conclusions

- Benefits of using deep neural network architectures for ranking emotional attributes
- Cross – corpora evaluations show that RankNet algorithms outperform RankSVM algorithms for $P@k$, τ
- Future Work
  - Use of other architectures (RNN-LSTMs) for preference learning to outperform DNNRegression
  - Ranking for emotional classes
  - Role of training data size in performance
    - Will we see better performance with increase in data size?

Thanks for your attention!

Questions?