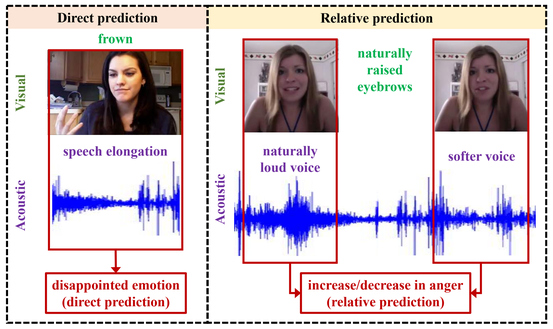
Multimodal Local-Global Ranking Fusion for Emotion Recognition

Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency
School of Computer Science, Carnegie Mellon University
{pliang,abagherz,morency}@cs.cmu.edu

Introduction



- Humans display their emotions through the **language, visual and acoustic modalities**.
- **Direct approach**: infer absolute emotions can be directly from **person-independent** behaviors.
- **Relative approach**: compare two video segments of the same person and determine the relative change in emotion intensities for **person-dependent** behaviors.
- Three easier subtasks:
  1. Local ranking of relative emotion intensities.
  2. Infer global relative emotion ranks.
  3. Incorporate both direct predictions from behaviors and relative emotion ranks from local-global rankings.

Related Work

- Non-temporal models.
- LSTMs and (Gated) Multi-view LSTM.
- Low-rank approximations to tensor products.
- Multiple fusion stages.
- Generative-discriminative objectives.
- Ranking algorithms for facial expression estimation.

Our approach:

1. Investigates a neural local-global ranking fusion approach for temporal multimodal data, allowing us to **incorporate probabilistic structure**.
2. Models both **speaker-independent** and **speaker-dependent features** via **direct and relative** predictions.
3. Uses **divide-and-conquer** to simplify the task.

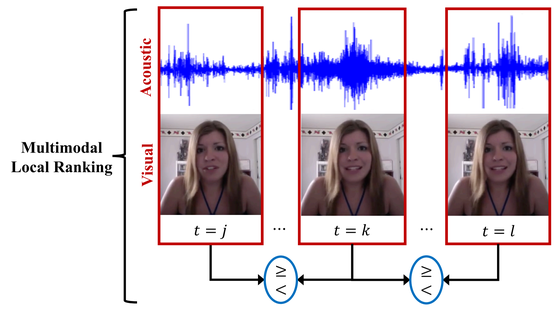
Notation

- Modalities $\mathcal{M}$, each a sequence with T time steps.
- Data from modality $m \in \mathcal{M}$ as $\mathbf{x}^m = \langle \mathbf{x}_1^m, \mathbf{x}_2^m, \cdots, \mathbf{x}_T^m \rangle$.
- $\mathbf{x}_w^m = \langle \mathbf{x}_{t-w}^m, \cdots \mathbf{x}_t^m, \cdots \mathbf{x}_{t+w}^m \rangle$ is the short video segment centered around time t with a time window w.
- Estimate sequence of emotion labels $\mathbf{y} = \langle y_1, y_2, \cdots, y_T \rangle$.
- Training dataset $\mathcal{D} = \{\mathbf{x}_{(i)}^m, \mathbf{y}_{(i)} : m \in \mathcal{M}\}_{i=1}^n$.

Multimodal Local-Global Ranking Fusion

1. **Multimodal local ranking**: a local ranking of emotion intensities between two short segments of a video.
2. **Global ranking**: use results of local rankings to infer global relative emotion ranks.
3. **Direct-relative fusion**: integrate direct emotion predictions estimated from observed multimodal behaviors with relative emotion ranks from local-global rankings.

Multimodal Local Ranking



1. Collect set of p pairs (J, K) by randomly sampling $1 \le j, k \le T$.
2. Construct multimodal local ranking dataset $\mathcal{D}_{local}$:

$$\mathcal{D}_{local} = \{\{\mathbf{x}_{j_w(i)}, \mathbf{x}_{k_w(i)}, r_{j,k(i)}\}_{j \in J_{(i)}, k \in K_{(i)}}\}_{i=1}^n \quad (1)$$

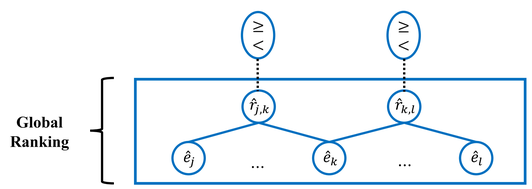
where $\mathbf{x}_{j_w} = \{\mathbf{x}_{j_w}^m : m \in \mathcal{M}\}$ is the video segment centered at j with window w and local rank $r_{j,k} = \mathbb{I}[y_j > y_k]$.

3. Use an LSTM on the differences of multimodal tensors:

$$\mathbf{x}_{local} = \bigoplus_{m \in \mathcal{M}} \mathbf{x}_{k_w}^m - \bigoplus_{m \in \mathcal{M}} \mathbf{x}_{j_w}^m \quad (2)$$

4. Classification layer on LSTM outputs to estimate ranks.

Global Ranking

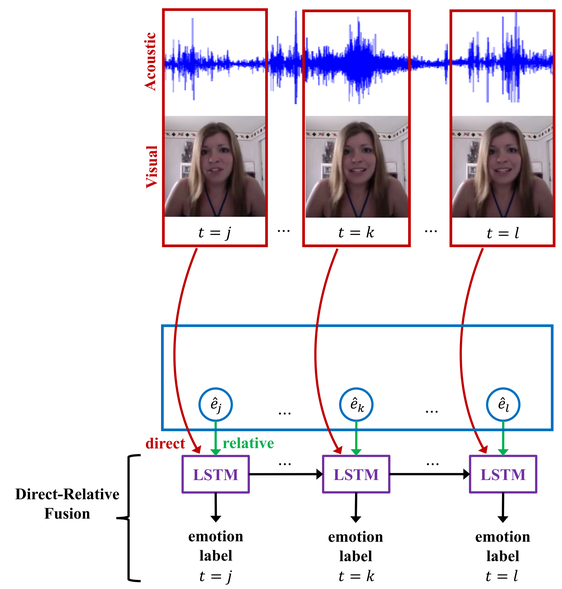


1. e_t are hidden variables that are not directly observed from the data while local rankings $r_{j,k}$ are observed.
2. Define prior over $e_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$.
3. Observe that

$$p(r_{j,k} = 1 | e_j, e_k) = p(e_j > e_k) \quad (3)$$

4. Given estimated ranks $\hat{r}_{j,k}$, we estimate global relative emotion ranks $\hat{e}_t$ using a ranking algorithm which involves message passing over factor graph models.

Direct-Relative Fusion



1. Use an LSTM on the concatenated multimodal data and global emotion ranks (Early Fusion LSTM):

$$\mathbf{x}_{fusion} = \left(\bigoplus_{m \in \mathcal{M}} \mathbf{x}^m\right) \bigoplus \hat{\mathbf{e}} \quad (4)$$

2. Regression layer on LSTM outputs to estimate emotions.
3. Flexibility: global emotion ranks can be incorporated using any multimodal fusion method.

Experiments

Dataset

- AVEC16 dataset (RECOLA).
- $T = 7501$ after alignment between the modalities, labeled continuously for arousal and valence.
- Metric: concordance correlation coefficient (CCC).

Baseline Models

1. EF-LSTM (Early Fusion), GF-LSTM (Gated Fusion), MV-LSTM (Multi-View) LSTMs.
2. Our model: **MLRF-p**, $w = 200$.

Results on Emotion Recognition

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
EF-(-/S/B/SB)LSTM	0.4327	0.4667
Gated-LSTM	0.3210	0.4667
MV-LSTM, view-specific	0.4530	0.4431
MV-LSTM, coupled	0.4300	0.4477
MV-LSTM, hybrid	0.4729	0.4924
MV-LSTM, fully connected	0.4293	0.4896
MLRF-500	0.4732	0.5063
MLRF-1000	**0.5049**	**0.5432**
Improvement over baselines	↑ **0.032**	↑ **0.0508**

- Incorporating local-global ranking estimates into simple models (EF-LSTM) > complex neural architectures.
- Sampling just 500-1000 local ranking pairs per video.

Discussion

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
MLRF-500 $w = 10$	0.4165	0.2377
MLRF-500 $w = 50$	0.4168	0.4175
MLRF-500 $w = 100$	0.4196	0.4340
MLRF-500 $w = 200$	**0.4732**	**0.5063**

- Increasing the window size w helps.

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
MLRF-500 direct predictions only	0.4327	0.4667
MLRF-500 relative predictions only	0.3646	0.0402
MLRF-500	**0.4732**	**0.5063**
MLRF-1000 direct predictions only	0.4327	0.4667
MLRF-1000 relative predictions only	0.4297	0.0846
MLRF-1000	**0.5049**	**0.5432**

- Sampling more local comparison pairs helps.
- Integrating both direct and relative approaches helps.

Conclusion

MLRF integrates **direct person-independent** and **relative person-dependent** perspectives via: 1) **multimodal local ranking**, 2) **global ranking** and 3) **direct-relative fusion**.

Acknowledgements

This material is based upon work partially supported by Samsung. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Samsung, and no official endorsement should be inferred.