RNN Architectures for Emotion Classification in Multimodal Emotion Recognition Systems

Deepan Das
John Ridley
July 19 2019

Advanced Methods for Emotion Recognition in Highly Automated Driving
Agenda

1. Background
2. Challenges
3. Existing Approaches
4. Demo
5. Proposed Solution
6. Conclusion
1.0 - Introduction

- **Autonomous driving is just being realised**
  - But there is a massive transition window
  - Awkward ‘monitored autonomy’ mix

- **Why monitor driver emotion?**
  - Safety
  - Comfort/Luxury

- **Continuous affective state monitoring**
  - Can they safely operate the vehicle?
  - Are they likely to be a hazard?
  - Can the vehicle influence/improve their mood?
1.1 - Let’s talk about our feelings

● Emotions are...
  ○ not inherently bad, but influence behaviour
  ○ classified in various ways
  ○ varyingly relevant to driving behaviour
  ○ temporally dynamic

Russell’s Circumplex Model

Plutchik’s Wheel of Emotions

Deepan Das - John Ridley
1.2 - Recurrent Neural Networks

- **Deep Neural Networks**
  - Fundamentally changed the ML landscape
  - Not directly applicable to temporally dependant data

- **Recurrent Neural Networks**
  - Unroll networks temporally
  - Cells propagate temporal context/state
  - Time dependent variable length sequences
  - Can also input/output single values
1.3 - RNN Zoo

Vanilla

LSTM

GRU

Vanishing Gradient

Gradient Pathways

Longer Term Memory

Bidirectional RNN

Stack RNN Cells

Reverse Gradient Flow

Vanishing Gradient Pathways

Longer Term Memory

Deepan Das - John Ridley
1.4 - Multimodal Emotion Classification

- Emotions observed in many ways - common sensor modalities

- Combine for ‘best of both worlds’
  - Certain modes can better indicate certain states [1]
  - More difficult than it seems
2.0 - Scope of Challenge

- **Which modes to utilise, given a driving context?**
  - Most physiological modes cannot be used - requires intrusive sensors
  - We focus on audio/visual modes (and variants) - scope of most existing research

- **How do we utilise RNN?**
  - RNNs work with sequence of aligned features
  - Where do we use the RNNs (before/after mode fusion)

- **Where/how are the modes and RNN(s) combined in the pipeline?**

2.1 - Mode-Based Challenges

- **Mode Variability**
  - Different sampling rates and numerical dimensions
  - Reliability/robustness of sensor or preprocessing methods

- **Mode Applicability**
  - Certain modes can better indicate certain states [1]
  - There are positive and negative conditions for both mode types

---

### Video

- **+** Driver clearly visible
- **-** Face (feature) not visible
- **-** Poor lighting
- **-** Driver not visible

### Audio

- **+** Driver conversing
- **-** Passenger conversing
- **-** Ambient noise
- **-** Silence
2.2 - Technical Challenges

- **Preprocessing**
  - Extraction of salient regions
  - Feature extraction (e.g. facial keypoints, audio features)

- **Fusion**
  - How/where are the modes combined (early or late)
  - How are mode failure states handled/trained

- **RNN Placement**
  - Before/after fusion (or both)
  - RNN type and depth
  - Combination with CNN
  - Resource and gradient limitations
3.0 - Problem Statement

- **Constraints for our reviewed approaches**
  - Audio/visual data for ‘emotions in the wild’
  - Discrete emotion classification (but also some regression techniques)

- **Some caveats**
  - There are numerous datasets all with different classifications and samples
  - No consistently used dataset - difficult to compare cross-paper results
  - ‘In the wild’ can imply acted or dramatised scenes
End-to-End Multimodal Emotion Recognition using Deep Neural Networks [1]

- Mode-wise custom CNN - Multimodal LSTM
- Output regressed arousal and valence
- Fusion approach yields best of both modalities
An Early Fusion Approach for Multimodal Emotion Recognition using Deep Recurrent Networks [2]

- Concatenated RNN - no mode dependent CNNs
- Input audio, eye, face, depth features (pre-labelleed & rate corrected)
- Classified 6 emotions
3.3 - Sun (2018)

Context-aware Cascade Attention-based RNN for Video Emotion Recognition [3]

- Image CNNs applied across video sequence with LSTM
- Face and context (whole) image modes

EXISTING APPROACHES

Deepan Das - John Ridley
Context-aware Cascade Attention-based RNN for Video Emotion Recognition [3]

- Proposed Context Attention mechanism
- Classified 8 emotions

Context-aware Attention-based RNN

Example of visual attention

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>55</td>
</tr>
<tr>
<td>Context</td>
<td>30</td>
</tr>
<tr>
<td>Concatenated</td>
<td>50</td>
</tr>
<tr>
<td>Parallel</td>
<td>45</td>
</tr>
<tr>
<td>Attention</td>
<td>40</td>
</tr>
</tbody>
</table>

Deepan Das - John Ridley
3.4 - Pei (2015)

Multimodal Dimension Affect Recognition using Deep Bidirectional LSTM RNNs [4]

- Deeper LSTMs (bidirectional)
- Combine mode-wise, multimodal RNNs and moving average
- Output regressed arousal, valence and dominance

**Diagram:**
- Audio
- Video Frames
- Deep Bidirectional LSTM (DBLSTM)

**Graph:**
- Accuracy (%)
- LR-MA, SYR, BLSTM, DBLSTM, DBLSTM-MA

Deepan Das - John Ridley
Multi-Feature Based Emotion Recognition for Video Clips [5]

- Late fusion, weighted by accuracy of each branch
- Not end-to-end trainable, to be avoided despite good performance

**Existing Approaches**

3.5 - Liu (2018)

- Face
  - Landmark Detector CNN
  - Stats on landmark distances
- Face
  - DenseNet, Inception
  - Stats on extracted features
- Face
  - Tuned VGG16
- Audio
  - Tuned SoundNet
  - LSTM

Late fusion, weighted by accuracy of each branch
Not end-to-end trainable, to be avoided despite good performance
3.6 - Guo (2018)


- Visual attention processed by LSTM as a set
- Skeleton data includes face, posture and hands

Example of visual attention, from [6]
Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction [7]

- Individual mode error modeled as measurement noise in Kalman
- Outputs continuous-time valence arousal values

3.7 - Brady (2016)

<table>
<thead>
<tr>
<th>Mode</th>
<th>Feature Extraction</th>
<th>Feature Extraction</th>
<th>Feature Extraction</th>
<th>Feature Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>SVM Regressor</td>
<td>LSTM</td>
<td>LSTM</td>
<td>LSTM</td>
</tr>
<tr>
<td>Video</td>
<td>CNN feature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRV</td>
<td></td>
<td>LSTM</td>
<td>LSTM</td>
<td>LSTM</td>
</tr>
<tr>
<td>EDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kalman Filter (Late Fusion)

- Outputs continuous-time valence arousal values

![Graph showing CCC values for different modalities](image-url)
4.0 - Demo Setup

Our Setup

- AFEW 2018 dataset [8]
- Focus on Neutral, Happy, Angry emotions
- Precomputed facial (from CNN of [9]) and audio features (from OpenSMILE)
- All code (except feature preprocessing) is our own
- RNN networks trained from scratch
- Tested various RNN architectures

Audio Descriptor ➔ RNN ➔ Output

Face Descriptor ➔ RNN ➔ Output

Validation Accuracy (%)

LSTM: 65%
B LSTM: 75%
GRU: 55%
B GRU: 60%

Deepan Das - John Ridley
4.1 - Demo Pipeline

Audio Descriptor

Face Descriptor

AFEW 2017
‘Emotions in the Wild’
Unseen Validation Set

Bidirectional LSTM

Trained by us

Output
4.2 - Our Demo
4.3 - Demo Performance

How does the model perform?

- Using our best-performing bidirectional LSTM model
- Trained with modalities both enabled and disabled
Drawbacks of existing approaches

- Noisy/missing audio to augment training makes more robust, but hurts performance
- Encoder-decoder with tied weights does not scale well
- Existing models not forced to discover correlations across modalities
- Different hidden units of existing models not forced to learn different modes

Multimodal Deep Learning [10]
Generating transcripts from both speech and ‘lip-reading’
How can cross-modal learning be combined with RNNs for emotion classification?

- **Train mode-wise networks to map to an ‘emotion space’**
  - Similar emotions map to similar positions in ‘emotion space’
  - Accomplished with a triplet loss and mode-wise encoder networks
  - Train on precomputed features and three emotions

- **RNN learns from ‘emotion space’**
  - Concatenate mode features like before, but after mapping into ‘emotion space’
  - RNN is trained after encoders
5.2 - Pipeline

PROPOSED METHOD

- **Audio Features**
  - Audio Feature Encoder

- **Video Features**
  - Video Feature Encoder

**Joint Embedding Space**

- **Triplet Loss**

**RNN**

- Option: FC
- OR
- Option: LSTM

Deepan Das - John Ridley
5.3 - Evaluation

- **Results**
  - Emotion classes embedded into ‘emotion space’
  - Projected to 2D using T-SNE approach - designed to show point distances
  - Embedding space not sufficiently separated

- **Possible Cause**
  - We use precomputed features - feature descriptors cannot change
  - Errors cannot propagate to descriptor networks to select separable features
5.4 - Advantages & Disadvantages

+ Modes have same meaning independent of each other
+ Easy to tell whether modes are in agreement
+ RNN no longer needs to learn how modes are related
+ Easy to add other modalities - just train another encoder
+ RNN can utilise the emotion description in the latent space

- Additional training steps & overhead
- Embedding difficult with precomputed features
- Projection to embedding space may remove mode-specificities
6.0 - Summary

- **Autonomous driving revolution - still a long way to go**
  - Cars need to monitor drivers to ensure safe ‘hybrid’ operation

- **Emotion recognition is a well established field**
  - But still very challenging - emotions difficult to classify consistently
  - Multimodal approaches provide measurable benefits

- **RNNs work well in Multimodal Emotion Recognition**
  - Take advantage of sequences of continuous features
6.1 - Approaches

- **Multimodal RNN Placement/Application Research**
  - Where does the RNN go? (Normally post fusion, but sometimes before too)
  - Where do we fuse features? (Later is generally better)
  - What type(s) of RNNs are used? (Bidirectional LSTM/GRU)

- **Cross-modal Approaches**
  - Generalise representation of concepts across different modes

- **Our Approach**
  - Combine both - unable to evaluate with pretrained features
6.2 - Future Direction

● How do we extract features from various modes?
  ○ Manage representations of emotions from different modes
  ○ Utilise HRV extracted from faces as an additional mode

● Is there a better way to combine features?
  ○ Deal with failed detections in certain modes
  ○ Get the cross modal representation working with descriptors

● How do we quickly and effectively train RNN Encoders?
  ○ More difficult than regular deep networks
  ○ Adopt cutting edge RNN encoder-decoder architectures
Questions


