Disentangled Adversarial Transfer Learning for Physiological Biosignals

EMBC 2020
Mo Han\textsuperscript{1}, Ozan Ozdenizci\textsuperscript{1}, Ye Wang\textsuperscript{2}, Toshiaki Koike-Akino\textsuperscript{2} and Deniz Erdogmus\textsuperscript{1}

\textsuperscript{1}Cognitive Systems Lab (CSL) - Northeastern University, Boston
\textsuperscript{2}Mitsubishi Electric Research Laboratories (MERL), Cambridge
1. Introduction

- **Physiological and Mental Status Monitoring**
  - Traditional method: electroencephalography (EEG) signal
    - surface (non-invasive) or implanted (invasive) electrodes
    - frequent calibration
  - Non-EEG physiological biosignals: temperature, heart rate, and arterial oxygen, etc.
    - wrist-worn platform
    - more effective, comfortable, and less expensive
  - Major issue: variability among different subjects or recording sessions

- **Transfer Learning**
  - Cope with the change in data distributions, in order to fit a wider range of users
  - Adversarial training
    - allow the representation to predict dependent variables
    - simultaneously taking advantage of an adaptive measure
    - control the extent of its dependency during training
1. Introduction

- Our work: adversarial inference approach
  - Exploit disentangled nuisance-robust representations
  - Trade-off between task-related features and person-discriminative information
  - Additional censoring network blocks: **Adversary block and Nuisance block**
    - jointly train the adversary, nuisance and classifier units
    - task-discriminative features are incorporated for unknow users dissimilar from training data
    - features from known subjects are projected to unknow but similar users’ data
  - Proposed disentangled adversarial transfer learning is applicable to other deep learning network approaches that are available
2. Methods: Disentangled Adversarial Transfer Learning

\{ (X_i, y_i, s_i) \}_{i=1}^{n} : \text{training dataset}

\[ X_i \in \mathbb{R}^{C \times T} : \text{raw data at trial } i \text{ recorded from } C \text{ dimensions for } T \text{ time samples} \]

\[ y_i \in \{0, 1, \ldots, L - 1\} : \text{label of user stress level status or task among } L \text{ categories} \]

\[ s_i \in \{1, \ldots, S - 1, S\} : \text{subject identification (ID) among } S \text{ individuals} \]
2. Methods: Disentangled Adversarial Transfer Learning

\[ z = g(X; \theta) \] : encoder, to learn the latent representation \( z \) from data \( X \)

\[ Z \] : latent feature, concatenation of \( z_a \) and \( z_n \) on a ratio of \( (1-r_N): r_N \)
2. Methods: Disentangled Adversarial Transfer Learning

![Diagram of Adversarial Transfer Learning]

**$z_a$:** input to the *adversary* network, aims to **conceal user-related information** $s$

Adversary: a classifier for user-related information $s$, with $\hat{S}_A$ as the output

$\Rightarrow$ let feature $z_a$ have a lower correlation on classifying $s$, i.e. **maximize** $\text{loss}_{\text{adversary}}(s, \hat{S}_A)$
2. Methods: Disentangled Adversarial Transfer Learning

\[ z_n \]: input to the *nuisance* network, aims to include user-related information \( s \)

Nuisance: a classifier for user-related information \( s \), with \( \hat{S}_N \) as the output

\( \rightarrow \) let feature \( z_n \) have a higher correlation on classifying \( s \), i.e. minimize \( \text{loss}_{\text{nuisance}}(s, \hat{S}_N) \)
2. Methods: Disentangled Adversarial Transfer Learning

$$\max_{\theta, \gamma, \psi} \min_{\phi} \mathbb{E} [\log q_\gamma (y | g(X; \theta), s) + \lambda_N \log q_\psi (s | z_n) - \lambda_A \log q_\phi (s | z_a)]$$

$$\min \text{loss}_{\text{classifier}}(y, \hat{y}) \quad \min \text{loss}_{\text{nuisance}}(s, \hat{s}_N) \quad \max \text{loss}_{\text{adversary}}(s, \hat{s}_A)$$
3. Experimental Evaluation and Results: Physiological Biosignal Dataset

- Dataset: physiological biosignal dataset for assessing human stress status levels
  - 4 stress status (L = 4):
    1. physical stress
    2. cognitive stress
    3. emotional stress
    4. relaxation
  - 20 healthy subjects (S = 20)
  - 7 channels (C = 7): biosensors containing
    1. electrodermal activity
    2. temperature
    3. heart rate
    4. arterial oxygen
    5. 6. acceleration
  - 300 time samples (T = 300): task of 5 minutes downsampling to 1 Hz
3. Experimental Evaluation and Results: Experiment Implementation

- **Parameters:**
  - o known: channel number $C = 7$, time sample $T = 300$, label number $L = 4$, subject number $S = 20$
  - o to be optimized: adversary regularization weight $\lambda_A$ and nuisance regularization weights $\lambda_N$
  - o to be optimized: nuisance representation rate $r_N$ among all features

- **Parameter optimization:**
  - o 1. first optimize $\lambda_A$ with only adversary block: $\lambda_A \in \{0.05, 0.1\}$ with $\lambda_N = 0$ and $r_N = 0$
  - o 2. fix the nuisance rate to $r_N = 0.2$: assume that the subject-related feature $z_N$ accounts for a small proportion among feature $z$ and keeps constant for all users and tasks
  - o 3. second optimize $\lambda_N$ with both adversary and nuisance blocks: $\lambda_N \in \{0.001, 0.005, 0.05, 0.01, 0.2\}$ with $r_N = 0.2$ and optimized $\lambda_A$ from step 1

- **Validation:** cross-subjects validation using a leave-one-subject-out approach
3. Experimental Evaluation and Results: Results and Discussion

<table>
<thead>
<tr>
<th>$\lambda_A$</th>
<th>$\lambda_N$</th>
<th>$\tau_N$</th>
<th>Main Classifier</th>
<th>Adversary Network</th>
<th>Nuisance Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Adversarial</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>79.88%</td>
<td>71.13%</td>
</tr>
<tr>
<td>Adversarial</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>79.97%</td>
<td>35.62%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>80.34%</td>
<td>8.08%</td>
</tr>
<tr>
<td>Disentangled Adversarial</td>
<td>0.1</td>
<td>0.001</td>
<td>0.2</td>
<td>80.62%</td>
<td>7.05%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.005</td>
<td>0.2</td>
<td>80.66%</td>
<td>7.90%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.05</td>
<td>0.2</td>
<td>80.04%</td>
<td>7.37%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>80.36%</td>
<td>8.08%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>80.22%</td>
<td>8.05%</td>
</tr>
</tbody>
</table>

- **Main classifier accuracy**: 4-class decoding of human stress
  - preferable: higher, indicates better discrimination of stress status levels

- **Adversary network accuracy**: 20-class decoding of subject ID
  - preferable: lower, indicates less subject-specific information are preserved in feature $z_a$

- **Nuisance network accuracy**: 20-class decoding of subject ID
  - preferable: higher, indicates more subject-specific information are preserved in feature $z_n$
3. Experimental Evaluation and Results: Results and Discussion

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_A$</th>
<th>$\lambda_N$</th>
<th>$\gamma_N$</th>
<th>Main Classifier</th>
<th>Adversary Network</th>
<th>Nuisance Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Adversarial</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>79.88%</td>
<td>71.13%</td>
<td>6.17%</td>
</tr>
<tr>
<td>Adversarial</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>79.97%</td>
<td>35.62%</td>
<td>6.15%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>80.34%</td>
<td>8.08%</td>
<td>6.20%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.001</td>
<td>0.2</td>
<td>80.62%</td>
<td>7.05%</td>
<td>39.03%</td>
</tr>
<tr>
<td></td>
<td>$\bf{0.1}$</td>
<td>$\bf{0.005}$</td>
<td>$\bf{0.2}$</td>
<td>$\bf{80.66%}$</td>
<td>$\bf{7.90%}$</td>
<td>$\bf{55.54%}$</td>
</tr>
<tr>
<td>Disentangled Adversarial</td>
<td>0.1</td>
<td>0.05</td>
<td>0.2</td>
<td>80.04%</td>
<td>7.37%</td>
<td>78.83%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>80.36%</td>
<td>8.08%</td>
<td>83.72%</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>80.22%</td>
<td>8.05%</td>
<td>87.26%</td>
</tr>
</tbody>
</table>

- Non-adversarial model: $\lambda_A = 0, \lambda_N = 0, \gamma_N = 0$
- Adversarial network: $\lambda_A = 0.1, \lambda_N = 0, \gamma_N = 0$
- Disentangled adversarial network: $\lambda_A = 0.1, \lambda_N = 0.005, \gamma_N = 0.2$
3. Experimental Evaluation and Results: Results and Discussion

- Non-adversarial model:
  \[ \lambda_A = 0, \lambda_N = 0, r_N = 0 \]
- Adversarial network:
  \[ \lambda_A = 0.1, \lambda_N = 0, r_N = 0 \]
- Disentangled adversarial network:
  \[ \lambda_A = 0.1, \lambda_N = 0.005, r_N = 0.2 \]
Thank you.