Review: Deep learning-based electroencephalography analysis: a systematic review. Roy, 2019

Dejan Markovikj       2/5/2020
BCI vs “classic” DL

• Nature of raw input – EEG picks up the electric potential differences on the scalp when post-synaptic potentials of the pyramidal neurons sum together
  – The potentials reflect neural activity

• Highly structured data
  – Usually 2D matrix (time and channel) of potentials

• Less developed as a field of research

• Data was not readily available and is more complex
  – There are more datasets available now
  – Non-stationary signal (poor generalization = more data)
  – Inter-subject vs. intra-subject variability
  – Harder to automate data preprocessing and annotation

• High temporal resolution
  – Effectively distinguishing between signals even a ms apart

• Low spatial resolution
  – Because of physical characteristics of the human brain and electrodes precision
  – High spatial correlation

• Low signal to noise ratio (SNR)
  – Low potentials can be interfered with environmental, psychological and activity-specific noise (artifacts)
DL for EEG – Pros & Cons

• Deep Learning pros
  – Better automation and less expert help
  – Potentially learning new features
  – Better generalization and more flexible application
  – New applications: generative models and transfer learning

• Deep Learning cons
  – Less data available
  – Data collection is expensive
  – Privacy and sensitivity issues as working with humans
  – Low SNR makes the problem harder
Goals and Approach

• Goals
  – Literature review on articles between 2010-2018 on applying DL to EEG
  – Extract trends and highlight interesting approaches
  – Formulate recommendations for future research

• Approach
  – Query relevant articles from major databases (PubMed, Google Scholar, arXiv)
  – Scan their references and add additional relevant articles
  – Extract various data items about data, preprocessing, architectures, results, reproducibility
  – Analyze extracted data, select papers and uncover trends
Data

Figure 3. Countries of first author affiliations.
Data contd.
Data contd.
Data contd.
Preprocessing

- **Data augmentation**
  - Attractive technique for EEG data
  - Overlapping window, adding gaussian noise or eye blink / muscle activity, using GAN, FT alterations, swapping electrodes, down-sample signal
  - Only 3 papers explored the importance of DA but 30+ used it
    - Improvement by up to 100% but not always

- **Data preprocessing**
  - Down-sampling, band-pass filtering, windowing, remove line-noise, interpolate bad channels
  - Majority of papers use at least one preprocessing step

- **Artefacts handling**
  - Usually require expert intervention
    - amplitude thresholding
    - identification and removal of high-variance segments
    - handling eye blinking and muscle activity noise
  - ½ of the papers didn’t handled artefacts
    - easier automation
    - less subjective process

- **Feature extraction**
  - Time-frequency features, statistical features, power spectral density
  - ½ use raw EEG (time series that has been pre-processed)
Architecture

- Almost ½ of the papers use CNNs
  - their popularity increases with time
  - Usually use raw EEG data

- Only 13% RNNs

- Usually unsupervised feature extraction then classifier on top

- Relatively shallow networks
  - Mostly up to 5 layers to 10 max 31
  - Better performance compared to going deep

- Sometimes temporal and spatial data is processed individually
  - EEG specific cause of the differences between the two
Training

- **Procedure**
  - Feature learning in unsupervised manner (standard)
  - ½ of the papers
  - Pre-trained models is used
  - ¼ of studies don’t discuss their training procedure!

- **Regularization**
  - To control the complexity and better generalization
  - L1/L2 weight decay, early stopping, dropout, label smoothing
  - More then half the paper don’t address regularization

- **Optimization**
  - Half of the papers don’t address this!
  - Most frequently used is Adam

- **Hyperparameter search**
  - Most popular grid search and Bayesian optimization
  - 80% don’t address this and those they do often use trial and error approach
Model Inspection

- Only 27% reported inspecting their models
- Most frequently analysis of the weights

<table>
<thead>
<tr>
<th>Model Inspection Techniques</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis of weights</td>
<td>[32, 41, 95–97, 120, 133, 148, 182, 189, 201, 220, 223, 230, 247]</td>
</tr>
<tr>
<td>Analysis of activations</td>
<td>[93, 97, 120, 172, 186, 214, 227, 231]</td>
</tr>
<tr>
<td>Input-perturbation network-prediction correlation maps</td>
<td>[17, 76, 166, 167, 211]</td>
</tr>
<tr>
<td>Generating input to maximize activation</td>
<td>[16, 158, 178, 207]</td>
</tr>
<tr>
<td>Occlusion of input</td>
<td>[33, 102, 194]</td>
</tr>
</tbody>
</table>
Reporting results

- Performance metric
  - Most popular accuracy

- Cross-validation
  - Almost half the studies didn’t use any cross-validation

- Subject handling
  - Inter-subject is most popular and growing

- Statistical testing
  - Only 20 studies used statistical testing
Reporting results contd.

- **Type of baseline**
  - Implementing standard models
    - More error prone
  - Comparing to published models
    - Not always available
  - 70% use at least one traditional model
  - 30+% at least one DL model
- **Data availability**
  - More than half the data is public
- **Code availability**
  - Mostly unavailable
- **Reproducibility**
  - 90% of the papers are hard or impossible to reproduce!
Discussion

- Key topic of DL for EEG
  - EEG classification
  - Automation of the process and discarding the traditional pipeline

- Data
  - How much data?
  - Data augmentation
  - More data = better models
  - Generalization among subject is harder problem than usual
  - Noisy data – we need better hardware
  - Raw EEG is better more often than not

- Architecture
  - RNN not as popular as CNN
  - GANs are getting more popular, RBMs and AEs less
  - Shallower networks are better

- Training
  - Not well explained in the studies
  - Optimization/regularization/hyperparameter tuning is rarely used
Discussion contd. + recommendation

- Training
  - Not well explained in the studies
  - Optimization/regularization/hyperparameter tuning is rarely used

- Reporting
  - Flawed metrics
  - Hard to compare

- Reproducibility
  - Hard to impossible

- Recommendations

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clearly describe the architecture</td>
</tr>
<tr>
<td>2</td>
<td>Clearly describe the data used</td>
</tr>
<tr>
<td>3</td>
<td>Use existing datasets</td>
</tr>
<tr>
<td>4</td>
<td>Include state-of-the-art baselines</td>
</tr>
<tr>
<td>5</td>
<td>Share internal recordings</td>
</tr>
<tr>
<td>6</td>
<td>Share reproducible code</td>
</tr>
</tbody>
</table>

Provide a table or figure clearly describing your model (e.g. see [33, 59, 167]).
Make sure the number of subjects, the number of examples, the data augmentation scheme, etc are clearly described. Use unambiguous terminology or define the terms used (for an example, see table 1). Whenever possible, compare model performance on public datasets. If focusing on a research question that has already been studied with traditional machine learning, clarify the improvements brought by using DL. Share code (including hyperparameter choices and model weights) that can easily be run on another computer, and potentially reused on new data.
Limitations

• Including arXiv preprints
  – Is it really bad?

• Missed articles
  – Balanced keywords not favoring particular architecture or task

• Focus on key points
  – Not considered things like weight initialization, data normalization etc

• Quickly changing field
  – Need for reviews more often to add relevant research advancement and shift in ideas
My take

- Problem: data availability
  - Transfer learning, data augmentation, pre-training
  - Push for creating baseline datasets like MNIST/ImageNet in Vision

- Problem: noisy data
  - Better hardware (and actually using the better hardware we have)
  - Proficiency in using the hardware and clear defined protocols
  - Techniques for noise reduction (estimating noise with GSVD, Wavelet decomposition and masking)

- Problem: result reporting and replicability
  - Following guidelines like the ones given in these two papers
  - Employing the same level of rigor as in other deep learning areas

- Problem: architecture and training
  - Clearly discussing and using optimization/regularization/hyper-parameter tuning etc
  - Why CNNs and why not deeper (probably data for the second)

- Problem: universal tools
  - It seems BCI is where DL was in 2010/2011 – need for more tools to easy comparisons and replicability

- Problem: features
  - No manually constructed features at any cost
Overview: CNN vs RNN

- **TCN** – temporal CNN that outperforms RNNs in a lot of cases

- **RNNs**
  - Parallelism is not possible
  - Fixed receptive field
  - Gradient can be unstable
  - Can use a lot of memory (particularly LSTM)
  - Variable length input

- **TCNs**
  - Parallelism is easy
  - Flexible receptive field size
  - Stable gradient
  - Low memory for training
  - Variable length input

- **Disadvantages of TCN**
  - More memory during evaluation
  - Transfer learning might be challenging

![Graph showing accuracy on the copy memory task for different lengths T. While TCN exhibits 100% accuracy for all sequence lengths, the LSTM and GRU degenerate to random guessing as T grows.](image)

<table>
<thead>
<tr>
<th>Sequence Modeling Task</th>
<th>Model Size ($\approx$)</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. MNIST (accuracy)</td>
<td>LSTM 70K, GRU 87.2, RNN 96.2, TCN 21.5</td>
<td>99.0</td>
</tr>
<tr>
<td>Permutated MNIST (accuracy)</td>
<td>LSTM 70K, GRU 85.7, RNN 87.3, TCN 25.3</td>
<td>97.2</td>
</tr>
<tr>
<td>Adding problem $T=600$ (loss)</td>
<td>LSTM 70K, GRU 0.164, RNN 5.3e-5, TCN 0.177</td>
<td>5.8e-5</td>
</tr>
<tr>
<td>Copy memory $T=1000$ (loss)</td>
<td>LSTM 16K, GRU 0.0204, RNN 0.0197, TCN 0.0202</td>
<td>3.5e-5</td>
</tr>
<tr>
<td>Music JSB Chorales (loss)</td>
<td>LSTM 300K, GRU 8.45, RNN 8.43, TCN 8.91</td>
<td>8.10</td>
</tr>
<tr>
<td>Music Nottingham (loss)</td>
<td>LSTM 1M, GRU 3.29, RNN 3.46, TCN 4.05</td>
<td>3.07</td>
</tr>
<tr>
<td>Word-level PTB (perplexity$^1$)</td>
<td>LSTM 13M, GRU 78.93, RNN 92.48, TCN 114.50</td>
<td>88.68</td>
</tr>
<tr>
<td>Word-level LAMBADA (perplexity)</td>
<td>LSTM -4186, GRU -14725, RNN -14725, TCN 1279</td>
<td>131</td>
</tr>
<tr>
<td>Char-level PTB (bpc$^1$)</td>
<td>LSTM 3M, GRU 1.36, RNN 1.37, TCN 1.48</td>
<td>1.31</td>
</tr>
<tr>
<td>Char-level text8 (bpc$^1$)</td>
<td>LSTM 5M, GRU 1.50, RNN 1.53, TCN 1.69</td>
<td>1.45</td>
</tr>
</tbody>
</table>
Thank you