CSE-5830: Project Proposal

Project Structure & Literature Review about CRF as RNN

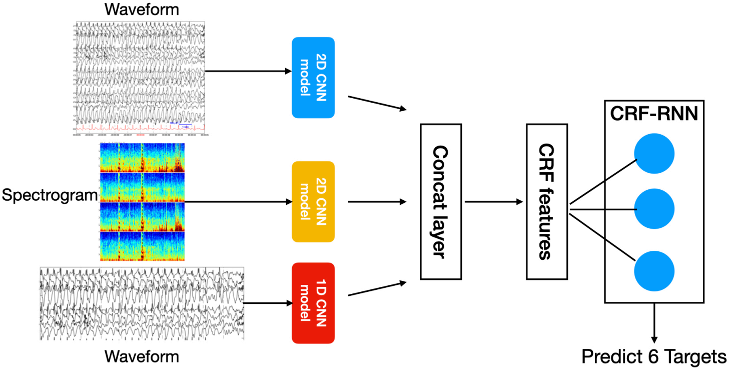
Xiaohang Ma

University of Connecticut
Department of Mathematics

UCONN
COLLEGE OF LIBERAL ARTS AND SCIENCES
DEPARTMENT OF MATHEMATICS

- Project Structure
  - Pipline of Our Model
  - Novelty of Our Approach
  - Team Roles
- End-to-end CRF Modeling with RNN
  - Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials
  - Conditional Random Fields as Recurrent Neural Networks
- Heterogeneity of Our Data

Model Structure



Model Structure

UCONN
COLLEGE OF LIBERAL
ARTS AND SCIENCES
DEPARTMENT OF MATHEMATICS

Novelty of Our Approach

1. Using CRF instead of *Softmax* function for final label predition.

Novelty of Our Approach

1. Using CRF instead of *Softmax* function for final label predition.
2. We will address the heterogeneous uncertainty of the label by introducing an environment invariant loss function.

Collaborative Efforts and Contributions

Topics related to this project:

- Statistics Methods [Xiaohui Yin]
  - Time series data analysis,
  - Heterogeneity data analysis.
- Probability Graphical Model [Shiying Xiao and Xiaohang Ma]
  - Build the CRF layer for label predition,
  - Mean field approximation inference and training of CRF model.
- Deep Learning methods [Shiying Xiao, Xiaohui Yin and Xiaohang Ma]
  - CNN and RNN layers for feature extraction
  - RNN-CRF layers for label discrimination

Manuscript Compilation and Editing [Shiying Xiao, Xiaohui Yin and Xiaohang Ma]

Project Presentation [Shiying Xiao, Xiaohui Yin and Xiaohang Ma]

Fully Connected Conditional Random Fields

A conditional random field $(\mathbf{I}, \mathbf{X})$ is characterized by a Gibbs distribution

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\left(- \sum_{c \in C_{\mathcal{G}}} \phi_c(\mathbf{X}_c|\mathbf{I}) \right),$$

where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph on $\mathbf{X}$ and each clique c in a set of cliques $C_{\mathcal{G}}$ in $\mathcal{G}$ induces a potential ϕ_c. The Gibbs energy of a labeling $\mathbf{x} \in \mathcal{L}^N$ is

$$E(\mathbf{x}|\mathbf{I}) = \sum_{c \in C_{\mathcal{G}}} \phi_c(\mathbf{x}_c|\mathbf{I}).$$

In the fully connected pairwise CRF model, $\mathcal{G}$ is the complete graph on $\mathbf{X}$ and $C_{\mathcal{G}}$ is the set of all unary and pairwise cliques. The corresponding Gibbs energy i :

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

Mean Field Approximation

- Instead of computing the exact distribution $P(\mathbf{X})$, the mean field approximation computes a distribution $Q(\mathbf{X})$.
- The approximation meature has two requirements
  - Q can be expressed product of independent marginals, $Q(\mathbf{X}) = \prod_i Q_i(X_i)$.
  - Q will minimize the KL-divergence $\mathcal{D}(Q \| P)$.
- This results the $\{Q(X_i)\}$ is the solution of the nonlinear systems:

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^{K} w^{(m)} \sum_{j \neq i} k^{(m)}(f_i, f_j) Q_j(l') \right\}.$$

Mean Field Approximation Inference

We will apply fixed-point interation to solve the above non linear systems numerically.

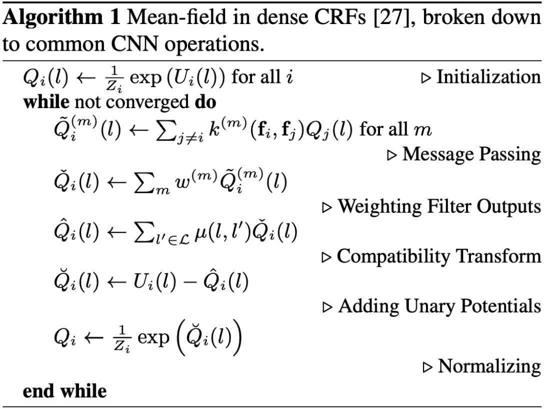
- Initialize Q by

$$Q_i(x_i) \leftarrow \frac{1}{Z_i} \exp\left\{-\phi_u(x_i)\right\}.$$

- $\hat{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(f_i, f_j) Q_j(l)$
- $\hat{Q}_i(x_i) \leftarrow \sum_{l' \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)} \hat{Q}_i^{(m)}(l)$
- $Q_i(x_i) \leftarrow \exp\left\{-\psi_u(x_i) - \hat{Q}_i(x_i)\right\}$
- normalize $Q_i(x_i)$
- repeat until convergence.

The time complexity of normalization factor $\{Z_i\}$ is $O(T \cdot N \cdot L)$, but the analytic solution needs $O(L^N)$!

Mean Field Iteration as Common CNN Operations



Algorithm 1 Mean-field in dense CRFs [27], broken down to common CNN operations.

```
$Q_i(l) \leftarrow \frac{1}{Z_i} \exp\left(U_i(l)\right)$ for all $i$    ▷ Initialization
while not converged do
    $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l)$ for all $m$
                                                    ▷ Message Passing
    $\check{Q}_i(l) \leftarrow \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$
                                                    ▷ Weighting Filter Outputs
    $\hat{Q}_i(l) \leftarrow \sum_{l' \in \mathcal{L}} \mu(l, l') \check{Q}_i(l)$
                                                    ▷ Compatibility Transform
    $\breve{Q}_i(l) \leftarrow U_i(l) - \hat{Q}_i(l)$
                                                    ▷ Adding Unary Potentials
    $Q_i \leftarrow \frac{1}{Z_i} \exp\left(\breve{Q}_i(l)\right)$
                                                    ▷ Normalizing
end while
```

A Plain RNN Architecture for End-to-end Training

Inefficient learning scheme of CRF

For the maximal log-likelihood training for CRF, the gradient w.p.t the partition function $\log Z(\mathbf{I};\theta)$ is intractable. Since

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log P(\mathbf{x}^{(n)}|\mathbf{I}^{(n)};\theta) = -\sum_{n=1}^{N}[E(\mathbf{x}|\mathbf{I}^{(n)};\theta) + \log Z(\mathbf{I}^{(n)};\theta)] \qquad (2.1)$$

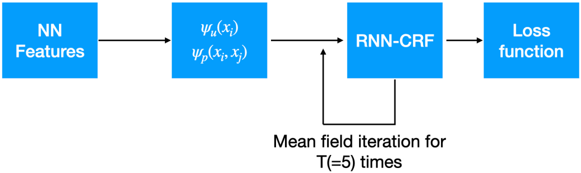
$$\Rightarrow \nabla_\theta \mathcal{L}(\theta) = -\sum_{n=1}^{N} \nabla_\theta [E(\mathbf{x}|\mathbf{I}^{(n)};\theta) + \log Z(\mathbf{I}^{(n)};\theta)]. \qquad (2.2)$$

However, a straightforward calculation leads to

$$\nabla_\theta \mathcal{L}(\theta) \log Z(\mathbf{I}^{(n)};\theta) = -\mathbb{E}_{\mathbf{x}^{(n)} \sim P(\mathbf{x}^{(n)}|\mathbf{I}^{(n)};\theta)} \nabla_\theta E(\mathbf{x}|\mathbf{I}^{(n)};\theta).$$

A Plain RNN Architecture for End-to-end Training

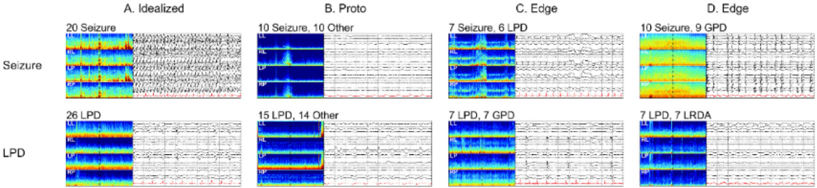
Mean field approximation the loss function



Heterogeneous Uncertainty

The EEG segments used in this competition have been annotated, or classified, by a group of experts. In some cases experts completely agree about the correct label. On other cases the experts disagree. We call segments where there are high levels of agreement “idealized” patterns. Cases where ~1/2 of experts give a label as “other” and ~1/2 give one of the remaining five labels, we call “proto patterns”. Cases where experts are approximately split between 2 of the 5 named patterns, we call “edge cases”.

Examples of EEG Patterns with Different Levels of Expert Agreement:



UCONN
COLLEGE OF LIBERAL ARTS AND SCIENCES
DEPARTMENT OF MATHEMATICS