Transformer for Alzheimer

Hui Wei

07/19/2019

Agenda

1. Task and Dataset
2. Dataset Preprocessing
3. Model description
4. Current Results
5. What is the next step

Task and Dataset

- **Task**: 1. Given a time series data, classify whether it is belonged to an Alzheimer patient or not. 2. Given a sliding window contained fixed number of visits, predict the cognitive status of the next visit. 3. Future: Given a history window and time gap, predict the cognitive status of the prediction window.
- **Dataset**: NACC dataset. Each patient has several visits. Demographic and biomarker data.
- **Statistics**: (1) 4,267 patients, 19,467 records (visits) in total (2) mean number of visits: 6.6, max number of visits: 14. (3) mean time span: 4.17 years, max time span: 13.19 years.

Preprocess the dataset

1. Get the label for each patient:

   Task 1: the label of the whole time series is the label of the last visit

   Task 2: the label of the sliding window is the label of the next visit

2. Divide the whole dataset into training, validation and test set

   Task 1: the ratio of alzheimer vs. no alzheimer in three sets are the same

   Task 2: randomly pick up patients into different sets

Preprocess the dataset

3. Drop the columns

Criteria for dropped columns:

(1) Texts (2) columns only having one value (3) administrative data (4) columns containing all the subjective diagnosis (5) columns defining the Alzheimer, like MMSE, CDR score, etc. (6) if there is some summary columns, drop all columns from which the summary columns can be derived. (7) columns having no descriptions in three documents. (8) columns containing years, like 2007, 1995. (9) columns whose number of records is less than 1000.

680 columns before we drop, 240 after they are dropped

Preprocess the dataset

4. Filling the missing data

Continuous data: fill with the median, two steps: (1) median of each patient (2) median of ALL patient

Categorical data: keep it as an another category

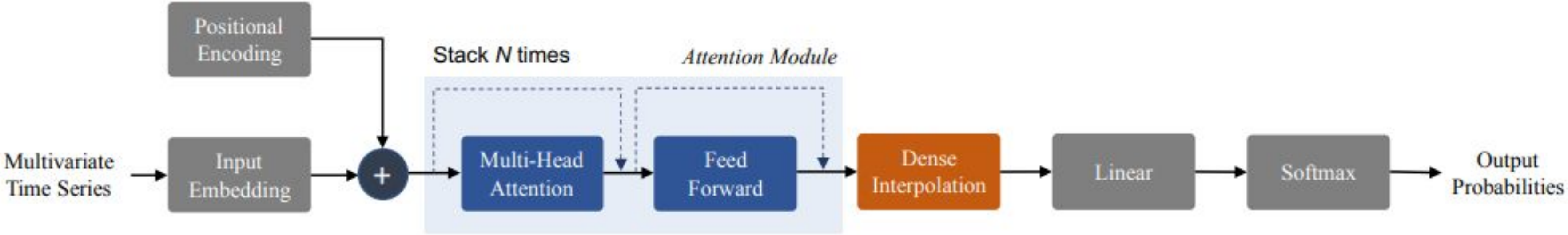
5. Feature Engineering

Continuous data: normalize them using the statistics of TRAINING set

Categorical data: encode as one-hot vector

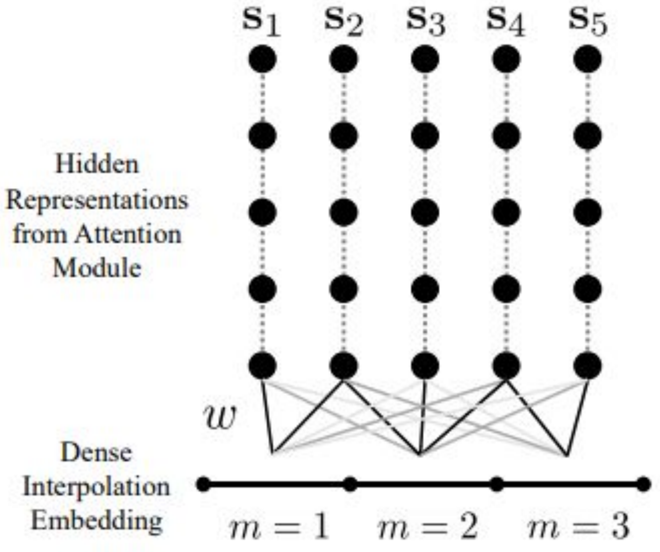
Feature: concatenation

Model Description



Song, Huan et al. “Attend and Diagnose: Clinical Time Series Analysis using Attention Models” *AAAI*,2018

Model Description



Dense Interpolation Embedding

Input : Steps t of the time series and length of the sequence T, embeddings at step t as $\mathbf{s}_t$, factor M.

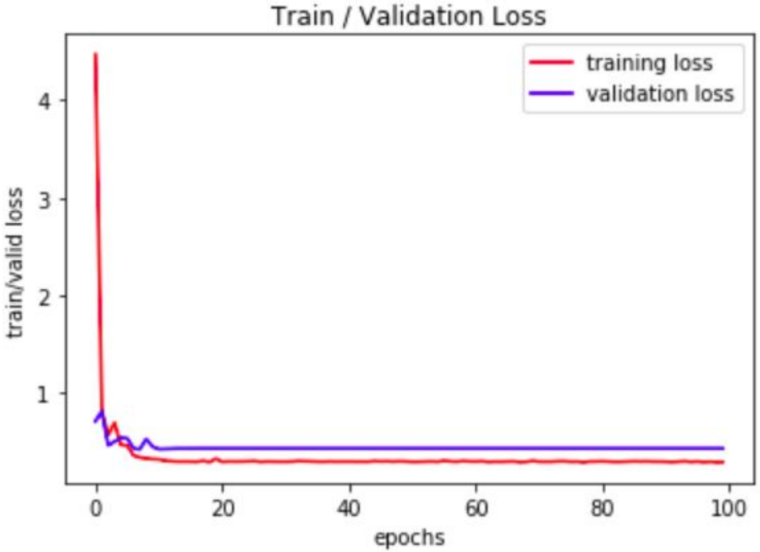
Output: Dense interpolated vector representation $\mathbf{u}$.

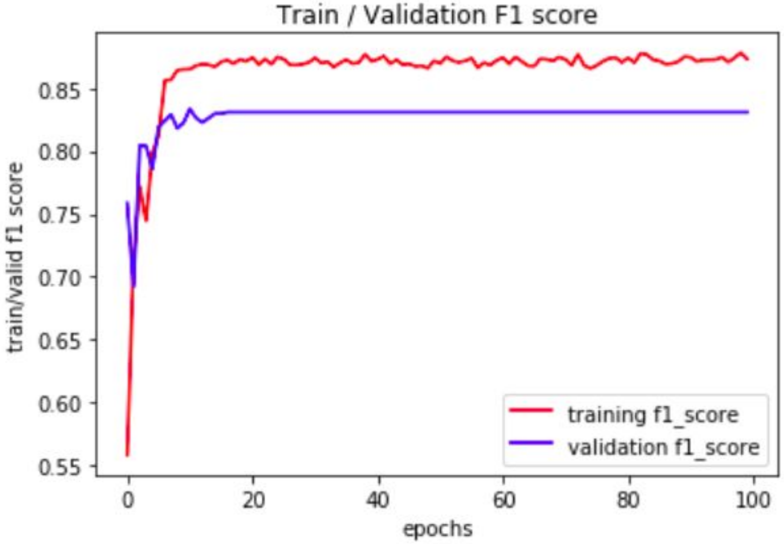
for $t = 1$ *to* T **do**
  $s = M * t/T$
  for $m = 1$ *to* M **do**
    $w = pow(1 - abs(s - m)/M, 2)$
    $\mathbf{u}_m = \mathbf{u}_m + w * \mathbf{s}_t$
  end
end

Song, Huan et al. “Attend and Diagnose: Clinical Time Series Analysis using Attention Models” *AAAI*,2018

Current Results

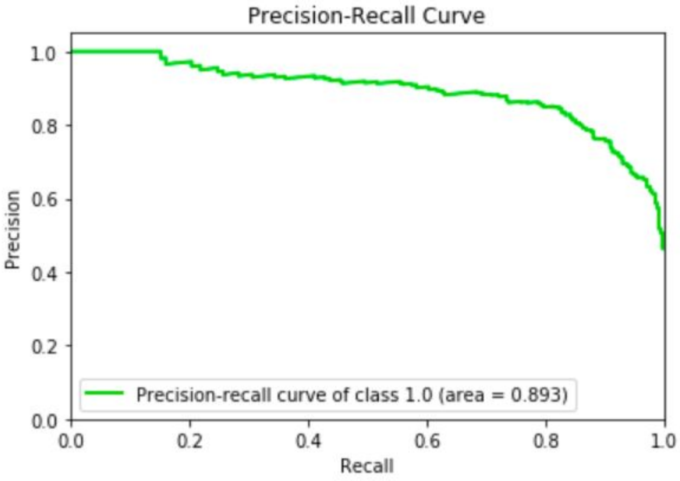
Classification:





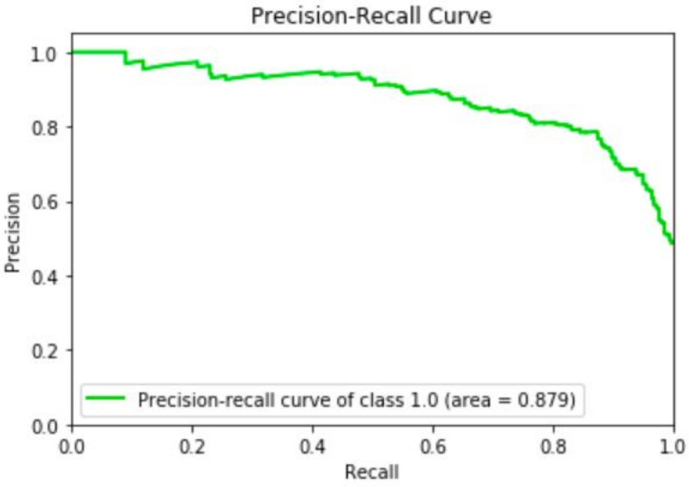
Current Results

Validation:



Best validation F1 score: 0.833575

Test:



Test F1 score: 0.812139

Next Steps

1. Task 2 trade-off: the number of visits of sliding window vs. training samples
2. Find the best number of visits in the sliding window
3. Use the pretrained model from Task 2 to fine-tune on the task 3

Thanks