

Rule induction for global explanation of neural classifiers

Madhumita Sushil, Simon Šuster, Walter Daelemans

Computational Linguistics and Psycholinguistics Research Center, University of Antwerp, Belgium
madhumita.sushil@uantwerpen.be

EXISTING APPROACHES

Word-level importance scores

No information about Interaction between multiple important words and corresponding class labels.

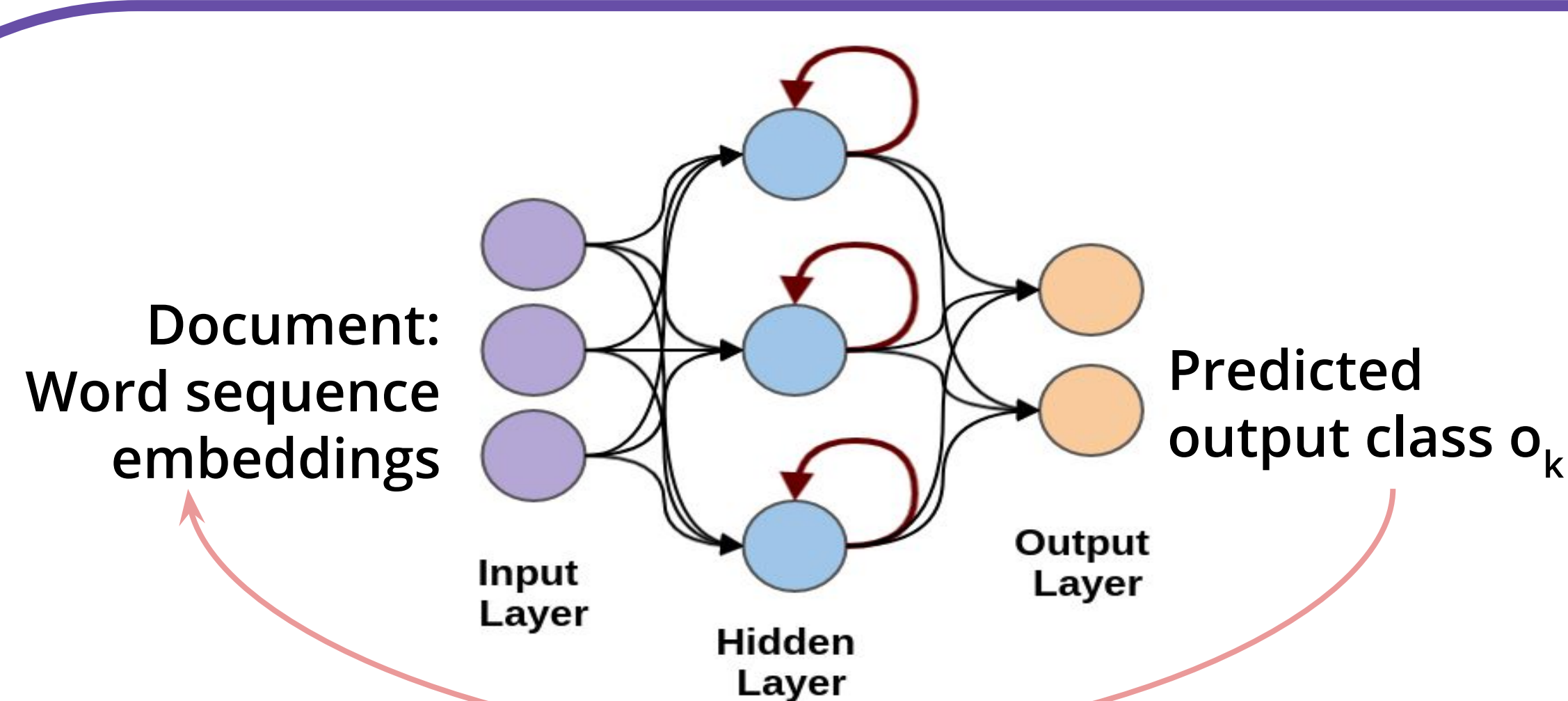
Explanation rules over original inputs

Don't encode knowledge about neural network parameters, and hence could learn completely different patterns despite the same outputs.

RESEARCH QUESTION

How can we induce rules that use neural network parameters to explain its decisions?

PROPOSED TECHNIQUE TO EXPLAIN RNNs



1. Input saliency, $G = \frac{\partial o_k}{\partial I}$

2. Compute word importance = $dot(I, G)$

3. Compute skipgram importance = $mean(word_imp)$

4. Retain the most important skipgrams

no signs of infection found. document1, class *non-septic*

infection is positive, found evidence. document2, class *septic*

5. Discretize skipgram importance

- +++ High positive impact on output probability
- ++ Low positive impact on output probability
- High negative impact on output probability
- Low negative impact on output probability
- ⊖ Absent in the input sequence

6. Rules as explanations

if no of infection is ++ and found is - then septic else: non-septic

SYNTHETIC DATASET FOR EVALUATION

Sentences sampled from MIMIC-III clinical corpus

- Containing an *infection_term*
- Containing a *measurement_term*
- Containing neither of the terms

Documents populated with 17 sentences each.

Gold labeling rule (using domain knowledge):

- If *infection_term* is not negated and min two *measurement_terms* are not negated:
 - Class *septic* 49%
 - Class *non-septic* otherwise

RESULTS - EXPLANATION ACCURACY %

	LSTM 100d, Emb 100d	LSTM 100d, Emb 50d	LSTM 50d, Emb 100d	LSTM 50d, Emb 50d
Classification	96.54	95.50	92.00	92.43
Baseline explanations*	76.10	78.17	83.89	84.96
Proposed method explanations	98.90	99.46	99.97	98.26

*Rules trained directly from the original input

RESULTS - EXAMPLE EXPLANATION RULES

hyperglycemia = ++ AND *to exclude* = ⊖ AND
evidence infection . = ⊖ AND *infection* = ++ AND
no infection . = ⊖ AND *no infection* = ⊖ AND
negative infection = ⊖ AND *or of infection* = ⊖ AND
fungal infection other = ⊖ AND *of infection in the* = ⊖ AND
altered = ++
 →septic (✓ 17466/17466)

tachypnea = ⊖ AND
meningitis = ⊖ AND
urinary tract = ⊖ AND
endocarditis = ⊖ AND
hyperglycemia = ⊖
 →non-septic (✓ 16015/16015)