Rule induction for global explanation of neural nets

Madhumita Sushil, Simon Šuster, Walter Daelemans
Computational Linguistics and Psycholinguistics Research Center, University of Antwerp, Belgium
madhumita.sushil@uantwerpen.be

EXISTING APPROACHES

Existing techniques for interpreting DNN in NLP:
- Input deletion
- Gradient computation
- Layerwise relevance propagation (LRP)
- Backpropagation using reference value (DeepLIFT)
- Learning explanation models: LIME, SHAP
- Attention weights

Open question:
How do we find the feature interactions that have been learned by a model for a certain class?

PROPOSED TECHNIQUE AND RESULTS: INDUCING RULES FOR INTERPRETING NEURAL NETS

BoW/
BoEntities, TF-IDF → Neural Classifier

Predicted output class ($o_k$)

Compute feature saliency, $G = \frac{\partial o_k}{\partial I}$

Reweight input features with their saliency scores in the network

Product of feature values and saliency scores

Select top reweighed features

Sensitivity Analysis / Mutual Information

Simplify the selected reweighed features to sign

Provides correlations b/w feature values and predicted output probabilities

+1: positive correlation (high feature value: high prob.)
-1: negative correlation (low feature value: high prob.)
0: absent feature value

Induce if-then-else rules using RIPPER-k on the correlation data to fit neural classifier predictions (one-vs-rest)

Feature-class associations that explain model predictions

Learned Rules (unordered):

- government = high and launch = absent and medical = absent and nasa = absent → Cryptography (✓ 45/46)

Text classification: medicine vs. space vs. electronics vs. cryptography (20 newsgroups data)

- Take blood pressure (treatment) = high and Nothing by mouth = absent and Coronary heart disease = high and Flagyl = absent → Diseases of the circulatory system (✓ 84/90)

Primary diagnostic category and in-hospital mortality prediction using EHR notes (MIMIC-III corpus)

- Dilantin = high and Thalamus, posterior lateral nucleus = high → Diseases of the nervous system (✓ 5/6)

- Pneumonia = high and Lung opacity = high and Non-specific ST-T changes by ECG = low and CT of pelvis w/o contrast = absent → Diseases of the respiratory system (✓ 7/7)

- Physical examination = high and Pregnancy with medical condition = high → Dies within hospital (✓ 221/222)

Results:
- Rule fidelity score to explain neural predictions:
  - 0.8 macro F1, high precision, lower recall.
  - Interdependence between features plays an important role for classification.