tRustNN: towards building trust in LSTM networks for an emotion recognition task, through data-driven, interpretable visualizations

Ioannis Chalkiadakis
Supervisor: Mike Chantler

Edinburgh Center for Robotics

ic14@hw.ac.uk

August 22, 2017

Contents

Motivation

- Advances in algorithms/hardware increase use of RNN, including safety-critical areas (automotive, social robotics, healthcare, privacy-related)
- Increase in demand for transparency
  - Engineers: better understanding, know model's limits hence improve system
  - Users: confident that system works as expected, trust $\rightarrow$ actually use it + quality rises
- DNN verification (theoretical/computer-aided): engineers
- Data-driven visualizations: engineers + non-expert users

Research gap

- Existing visualizations of RNN internals inform about training process
- Limited work on explaining what/how the models learn in an *interpretable??* way
- Ambiguous, vague definition
- Clarify notion $\rightarrow$ identify common guidelines on creating interpretable models $\rightarrow$ improve and *use* them

Research question

- Based on [1] expect: RNN experience same frailties, how can we identify them and show them to user
- How can we employ visualization methods to interpret a recurrent neural network for sentiment analysis, thus help novice users understand it, and provide advanced users with hints towards improving their network?

Project goals

- Initially
  - Generate adversarial inputs for an LSTM
  - Find visualization methods that allow end-users to investigate and understand the LSTM and its failures
- Re-plan
  - Framework to visualize LSTM in an interpretable way for non-experts
  - Provide a critical evaluation of the framework and the benefits we expect (+pilot)
  - Provide initial ideas for full experiments

Project outcome

- Framework
  - Left plot: parameter space
  - Wordcloud: raw input correlated with LSTM information in an interpretable way
  - Right: control panel
  - Bottom: visualization of LSTM attention span + raw input
- User can load a trained model + internals (LSTM hidden states + cell states)
- Code to extract required information for the visualizations $\rightarrow$ explore different LSTM models

Evaluation and outlook to the future

- Gains: thorough introduction to ML interpretability, experience with LSTM/user-based studies for future
- Contributions: framework for basic LSTM exploration, LRP with fully-conn + embedding layer
- Future roadmap
  - carry out experiment as planned
  - interpretability in ML in robotics applications: explain system's (sequence of) actions to users/collaborating robots $\rightarrow$ improve action planning, improve domain knowledge from robots' explanation

Key Reference


Anh Nguyen, Jason Yosinski, and Jeff Clune,
"Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,"
in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.

Backup

- KMeans on parameter values does not seem to be helpful under current projection (explainable PCA?)
- DBSCAN on all reviews/single review: ladder of abstraction, interpretable clusters of neurons?, DBSCAN (high-density clusters): segmented neuron space
- Positive-negative clustering: are there positive-only or negative-only neurons?, based on LRP of LSTM-layer, if decision==neg: positive neurons expected to have green or light red highlight
- Internal clustering: state identification?
- Click on neuron: highlight words according to neuron's LRP value, how important was that neuron to wc words?
- Network focus - LRP plot: short attention span