Comparing Attention-based Convolutional and Recurrent Neural Networks: Success and Limitations in Machine Reading Comprehension

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu and Ngoc Thang Vu
Institute for Natural Language Processing (IMS)

Research Agenda
- Analyze state-of-the-art neural-based models for reading comprehension
- Comparison with human performance under adversarial attacks
- Study human strategies to develop improved machine reading comprehension models
- Data: MovieQA dataset [Tapaswi et al., 2016]:
  - Crowd-sourced multiple-choice questions with five answer candidates on 408 movie plots
  - 9,848 training, 1,958 development (dev), 3,128 test questions

Take-Home
- New model with state-of-the-art results on MovieQA
- RNN-LSTM models outperform CNN models
- But adversarial attacks indicate they might suffer from the same weaknesses
- Models seem to learn matching patterns rather than performing plausible inferences as humans do
- Promising direction: integrate entailment, answer elimination, coreference resolution

Case Study: Human Strategies
Human annotator solves questions that were difficult for the models and analyzes own strategies:

Textual entailment: 60% of cases

P: ... They seize control of a Manhattan bank and take the employees and patrons hostage ...

Q: Do the robbers take people in the bank hostage?

A: 1) Never 2) No, they don’t 3) Yes, they do 4) No, they don’t, because there are many women and children 5) They take only the employee who has access to the vault, as they don’t want to hurt anyone

Elimination and heuristics: 44% of cases

P: ... In September 1965, on a New England island called New Penzance, 12-year-old orphan Sam Shakusky is attending Camp Ivanhoe ...

Q: Where is New Penzance located?

A: Off the coast of ... 1) Washington state 2) North Carolina 3) Virginia 4) Florida 5) California

Coreference resolution: 36% of cases

P: ... After the heist, Stigman follows orders to betray Trench and escape with the money, managing to pull his gun right as Trench is about to pull his own ...

Q: What does Stigman do with the money?

A: 1) He splits it with Trench 2) He takes it 3) He turns it over to the authorities 4) He leaves it in the vault 5) He leaves it for Trench

Hierarchical Compare-Aggregate Model

MovieQA accuracies:
- system: AddQ + all question words
  - [Wang and Jiang, 2016] 72.9
  - [Liu et al., 2017] 79.0
  - [Dzendzik et al., 2017] 80.0
- our CNN ensemble 82.6
- our RNN-LSTM ensemble 84.4
- our CNN RNN-LSTM ensemble 84.8

State of the art!
See http://movieqa.cs.toronto.edu/leaderboard/

Black-Box Adversarial Attacks
Access to outputs: labels and confidence scores

Word Level: Lexical Substitution
Manually define meaning-preserving substitutions for most frequent development set question words, e.g., friend → buddy

Accuracy on development set
- CNN ensemble: 81.7
- RNN-LSTM ensemble: 83.8
- CNN RNN-LSTM ensemble: 84.3

Sentence Level: Distracting Sentence
Add nonsense sentence to plot [Jia and Liang, 2017], e.g. What aziz what do do what clothing opens do do

Worst words in distracting sentence come from:
- AddC 1000 most frequent words in Brown corpus
- AddQA AddQ + incorrect answer candidates words

Word-level and sentence-level attention weights

White-Box Adversarial Attacks
Access to word- and sentence-level attention weights

Word Level: Replace attended Words
Replace k words with highest attention by random words in most attended sentence

Sentences

Hierarchical extension of [Wang and Jiang, 2016]:
Compute confidence score \( c_j \) of answer candidate \( A_j \) given question \( Q \) and plot sentences \( P_i \ldots P_n \):

Aggregation:
- Convolutional neural network (CNN): 1D convolution, filter sizes (1,3,5)
- Recurrent neural network with long short-term memory (RNN-LSTM): Max pooling over hidden states of single-layer unidirectional LSTM

Open Source Code & Contact

github.com/DigitalPhonetics/reading-comprehension
thang.vu@ims.uni-stuttgart.de