Transfer Learning for QA Systems

Kenta Takatsu, Chetan Valivela, Yuji Akimoto

February 26, 2018

1 Motivation & Potential Impact

A system that is able to acquire "knowledge" and answer questions about that knowledge is a significant step towards general artificial intelligence. As a long term goal, we aim to have a system that can acquire information from resources such as syllabi and textbooks from Cornell classes and then answer questions about them. We envision that this will automate the answering of simple and/or repetitive Piazza questions, leading to faster responses and allowing course staff to focus on more important questions. As a shorter term goal for the semester, we aim to create a system that answers question from SQuAD (Stanford Question Answering Dataset) [4] - we believe that this project is feasible as the SQuAD dataset limits the domain to the set of questions where the answer is a pointer to a portion of a provided piece of text.

2 Goals & Deliverables

Long term

- Make a system that can read a syllabus/textbook and answer questions about it.
- Build a system such that the models can generalize a "learning method" instead of memorizing domain specific knowledge. This way, our system can perform on the text corpus from completely foreign topics.

Short term

- This semester: create system to answer SQuAD dataset questions.
  1. Identify the sentence(s) that are most similar in topic to the question.
     - Start with a simple nearest neighbor algorithm with cosine similarity
     - Implement a transfer learning algorithm to optimize the function g that gives rise to the weighted similarity (i.e. words/topics that appear often globally should count less when computing similarity)
     - Finally, implement a memory augmented system to adjust mapping function g for different input features
  2. Given that sentence, find the answer.
     - Identify PoS type of answer (noun, proper noun, time, place, etc...)
     - Narrow down potential candidate answers: eliminate words shared by both question and relevant sentence (including synonyms, conjugated verbs, etc..)
     - Use TF-iDF embedding combined with a logistic regression as our baseline
     - Increase the complexity of the model as we need
  3. Assess the accuracy of our model by combining 2 methods above
     - Measure the accuracy for every improvements we make and identify the bottleneck

3 Responsibilities & Workload

Common language: Python. Plan to primarily use PyTorch (will involve some learning for Kenta, Yuji) and TensorFlow.

Very regular meetings to avoid procrastination and keep up progress.

- Monday after weekly meeting
- Thursday afternoon
- Saturday afternoon
- Sunday (if needed)

4 Schedule

2 completely separate tasks: finding relevant sentence, finding relevant words within sentence.

- Recreate baseline model
  - Filtering potentional candidates by kNN and cosine similarity
  - Finding the answer with tfidf and logistic regression
- Embed sentences as vectors
- Verify (manually) that it is reasonable
  - Instead of measuring the accuracy of the final outputs, we are going to test our 2 stages, filtering and searching, separately.
- Explore PoS tagging. Can we reliably identify proper nouns, times, places, verbs etc.
- Weighted sentence similarity - figure out which terms are key to the sentence, which sentences are specifically relevant.
- Label training data answers as PoS.

5 Datasets

- SQuAD [4]
- Course Syllabus / Textbooks
  - Collect data from the current members of CDS
  - Manually curate and construct the labeled data for the future

References

[1] Chuong B. DO, Andrew Ng. *Transfer Learning for Text Classification.* Stanford University

[2] Adam Santoro, Sergey Bartunov, Matthew Botvinick *One-shot Learning with Memory-Augmented Neural Networks.* Google DeepMind

[3] Tushar Semwal, Gaurav Mathur, Promod Yenigalla, Shivashankar B. Nair *A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks.* Indian Institute of Technology Guwahati

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016; arXiv:1606.05250.