DATASCI 207

Nedelina Teneva, PhD
nteneva@berkeley.edu

School of Information, UC Berkeley

Announcements

- **Exam:** just to test your knowledge, not used for assessment (yet)
  - Need to log into Google with your berkeley.edu email to access
- **HW10 is due Sunday (final homework this semester)**
  - Don't worry about exact numbers for Exercise 4. The idea is to do an ablation study.
- **Fairness Reading Material for next week**

  - Watch Margaret Mitchell's talk Keynote talk at the Stanford Center for Research on Foundation Models Workshop (at least the first ~30 mins)

  - How big data is unfair.

  - Machine Bias

  - Fairness in ML Course

Announcements - Project Details

- **Project Submission Guidelines and Grading**
  - *In your presentation include the names of each group member. Explain their contribution to each of the 5 components below – e.g, X did data processing, Y implemented component linear regression experiments, etc. Cite any outside code/materials used.*
  - Grading for final project (code + presentation):
    - **(20%) Code and presentation submission**: organize and document your code. You can create a repo with the code and presentation and add me to it (*username: methanet*)
    - **(20%) Dataset and EDA:** Describe in detail the data that you are using, including the source(s) of the data, and relevant statistics and EDA.
    - **(30%) Approach and Models:** Describe in detail the hypothesis tested and the models used. Include tables, figures, graphs to illustrate your findings and conclusions.
    - **(20%) Tuning and Improvements:** Include any hyper parameter tuning and/or any ensemble/boosting approach used for improving the results from your selected models.
    - **(10%) Conclusions and Checklist:** Summarize the key results and possible future work. Follow the NeurIPS checklist and answer all the questions (on a separate slide).

Last Week's Recap

- Sequential modelling
- Embeddings
- CNN for 1D data
- Application: Sentiment analysis based on drug reviews
    - Demo exercise: https://github.com/MIDS-W207/nteneva/blob/main/live_sessions_current/week10/CNN1D.ipynb

What did we learn so far?

- Learn how to formulate a learning problem (objective function, loss, optimization, performance evaluation, hyper parameter optimization)
- Analyzed different models (supervised & unsupervised) and experimented with different data modalities (tabular data, text, images)
    - Examples: logistic regression, trees, GMMS, neural networks (Feed forward, CNN), k-nn….

Today’s Objectives

- Discuss “ingredients” for ML success in practice
- Debugging learning curves
- End-to-end ML using Question Answering/Voice search application as a motivating example
- Short intro to RNN/LSTM

ML in practice – the necessary recipe “ingredients”

- Data
  - Is my data high quality?
  - Is the data biased?
  - Do I have data? If not what data do I need to create/annotate?
- Algorithms/Models
  - Baselines – usually simple, goal is to produce a POC to test feasibility, not to get the most optimal result
  - Incrementally produce more complex models by varying data and model complexity
- Compute & Storage
  - Models may be trained locally but the inference always needs to be run in production
  - Need to consider scalability and provenance when retrieving the data from various databases to use for ML model training
  - more on this in the ML@Scale course
- Model Performance and Downstream impact
  - My algorithm has satisfactory precision/recall/accuracy - does it have a positive impact on downstream KPIs
    - e.g., does my improved review sentiment classifier lead to better customer satisfaction with their purchases?

Motivating Example – Question Answering

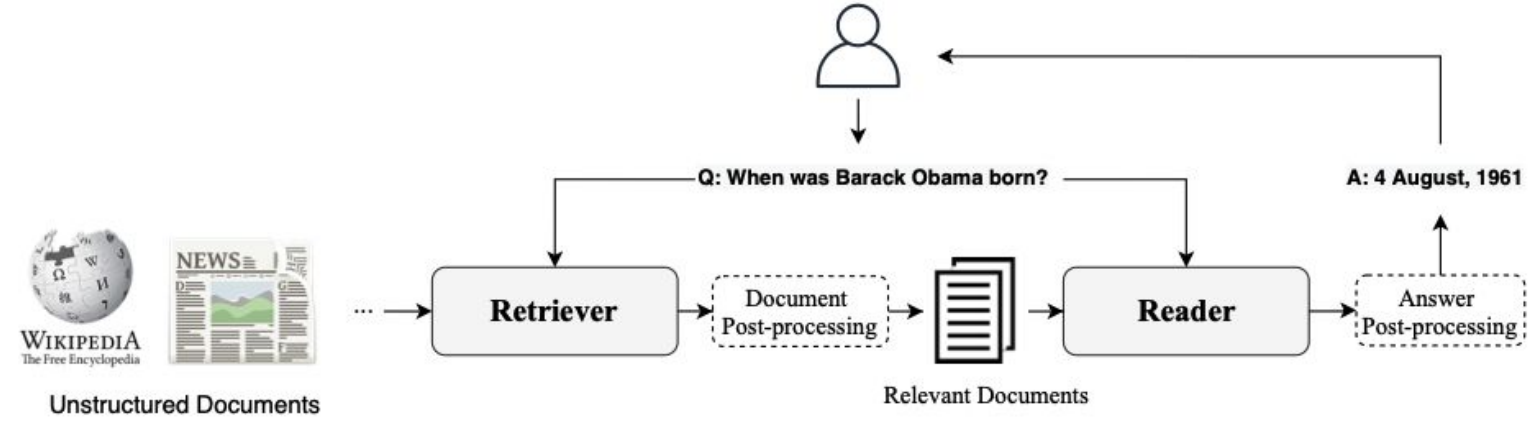


Fig. 3: An illustration of "Retriever-Reader" architecture of OpenQA system. The modules marked with dash lines are auxiliary.

Source: https://arxiv.org/pdf/2101.00774.pdf

Neural QA

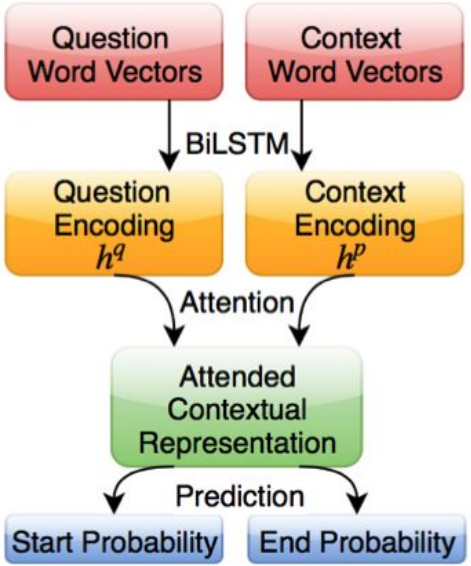
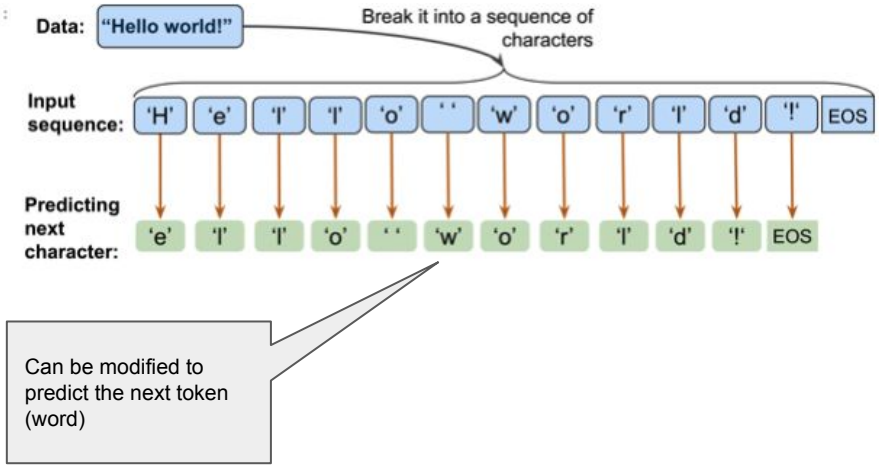
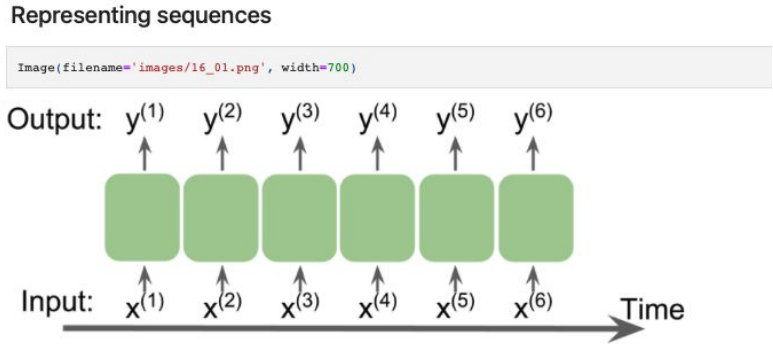


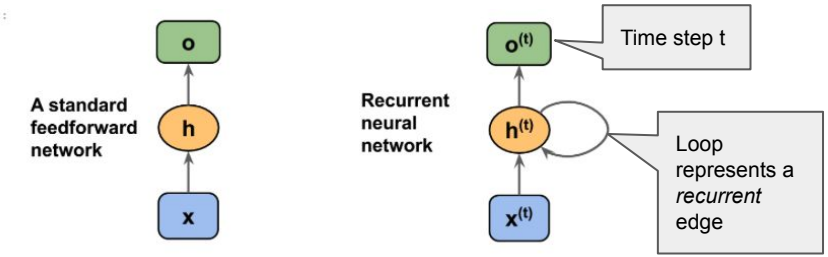
Figure 1: Common architecture of neural QA models.

Source: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761224.pdf

RNN/LSTM





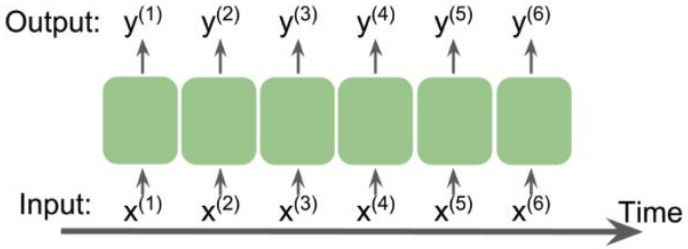


Demo (RM Chap 16): [Ch16_part1.ipynb]

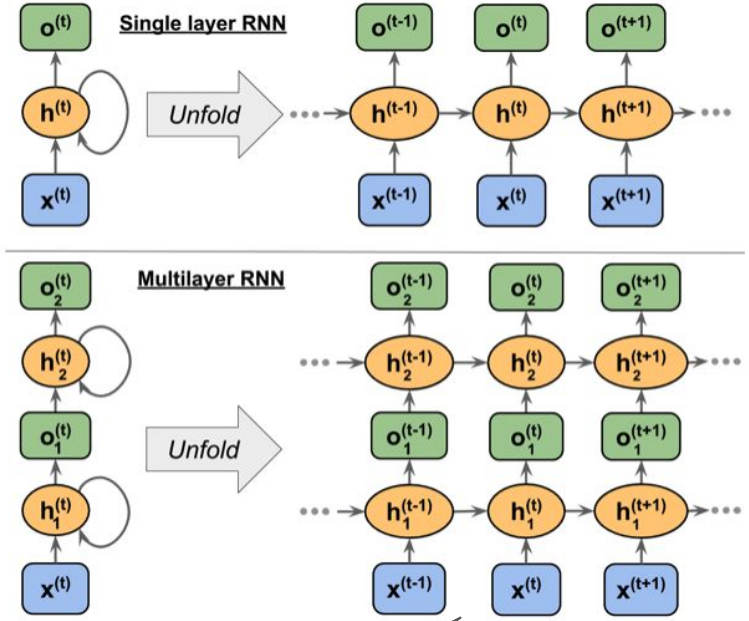
RNN/LSTM

Representing sequences

```
Image(filename='images/16_01.png', width=700)
```



[6]:



many-to-many architecture

QA Demo

- https://github.com/prashil2792/Question-Answering-System-Deep-Learning
- Uses the Facebook babi dataset https://github.com/facebookarchive/bAbI-tasks (answer questions based on stories)