Presentation 2:
Long Document Summarization

CS4624 Multimedia/Hypertext/Information Access

Team: Junjie Cheng

Instructor: Edward A. Fox

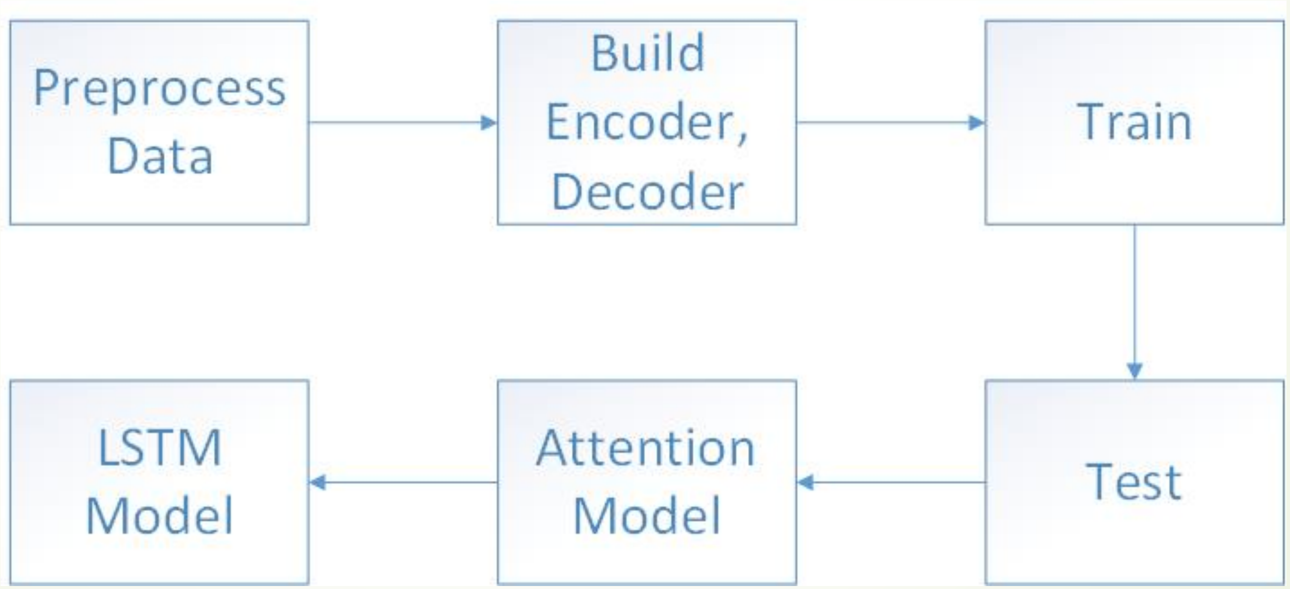
Virginia Tech, Blacksburg VA 24061, Mar 13, 2018

Outline

- Project Design
- Current Status
  - Data Preprocess
  - Encoder/Decoder
  - Training
  - Testing
- Future Plan
- Acknowledgements and References

Project Design

- **Purpose**: Generate abstract from long document by deep learning.



Preprocess Data

Build Encoder, Decoder

Train

Test

Attention Model

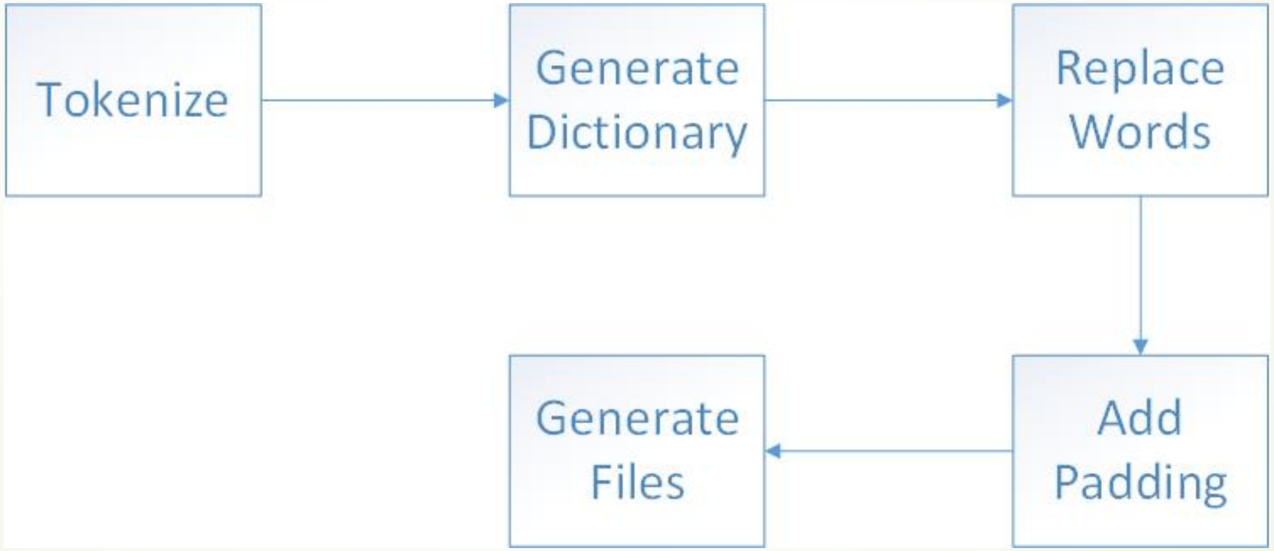
LSTM Model

Current Status

- Data Preprocess
- Encoder/Decoder
- Training
- Testing

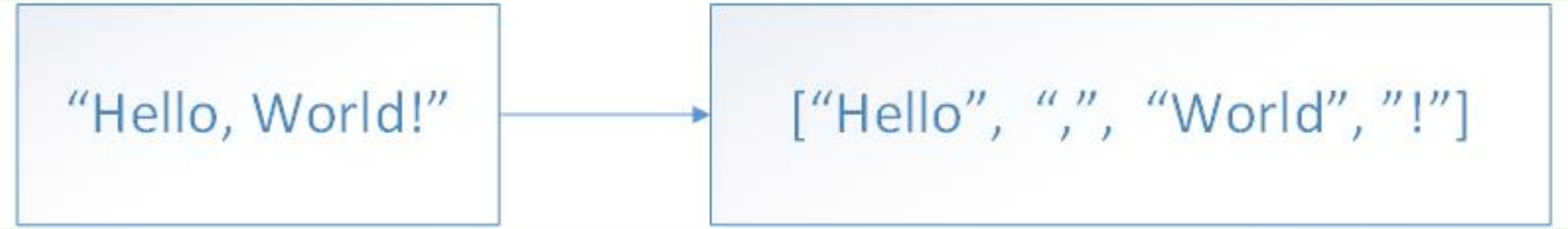
Data Preprocess

- 300,000 articles from CNN/Dailymail



Data Preprocess: Tokenize

- Tokenize words, symbols and numbers



“Hello, World!”

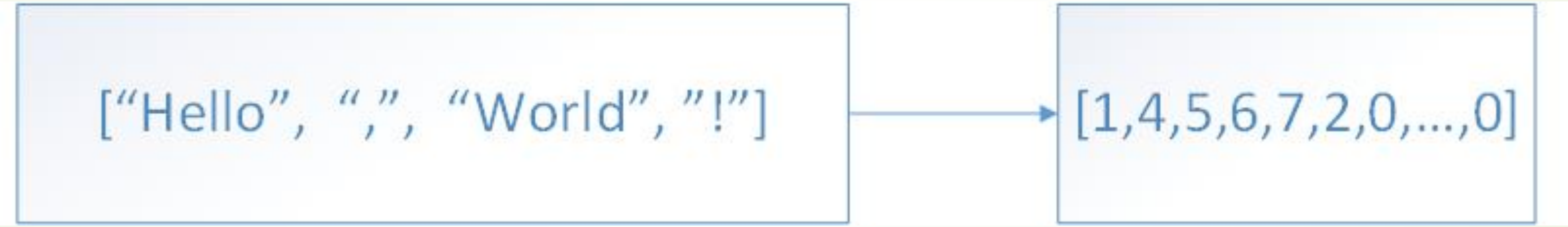
[“Hello”, “,”, “World”, ”!”]

Data Preprocess: Generate Dictionary

- Generate a dictionary includes all words and symbols
- For example:
  - “Hello” → 4
  - “,” → 5
- Also include:
  - “<Padding>” → 0
  - “<Start of Sentence>” → 1
  - “<End of Sentence>” → 2
  - ”<Unknown Word>” → 3

Data Preprocess:
Replace Words and Add Padding

- Replace words by ID
- Add <SOS>, <EOS>, and <Padding>

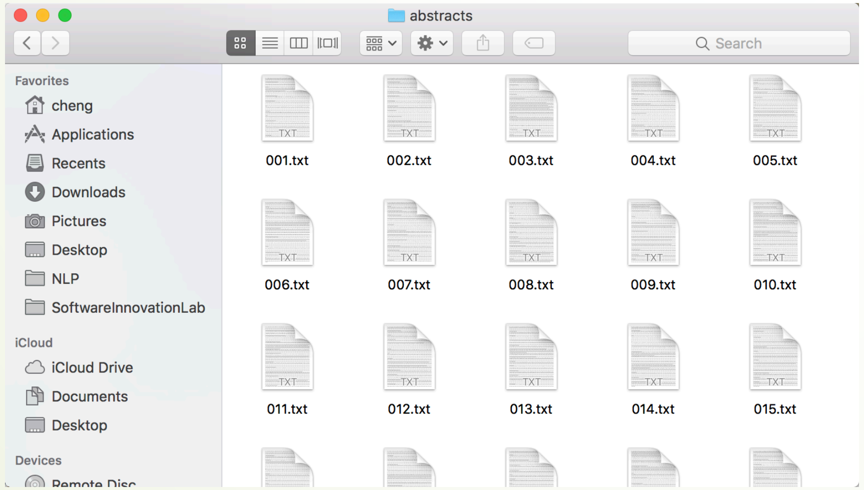


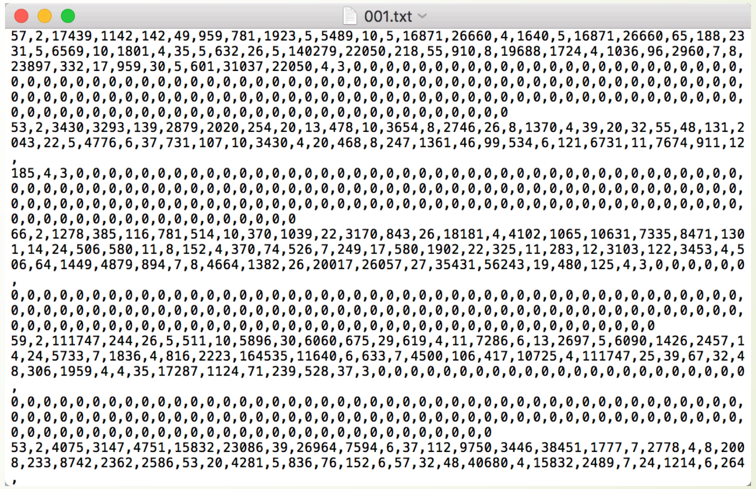
[“Hello”, “,”, “World”, ”!”]

[1,4,5,6,7,2,0,…,0]

Data Preprocess: Generate Files

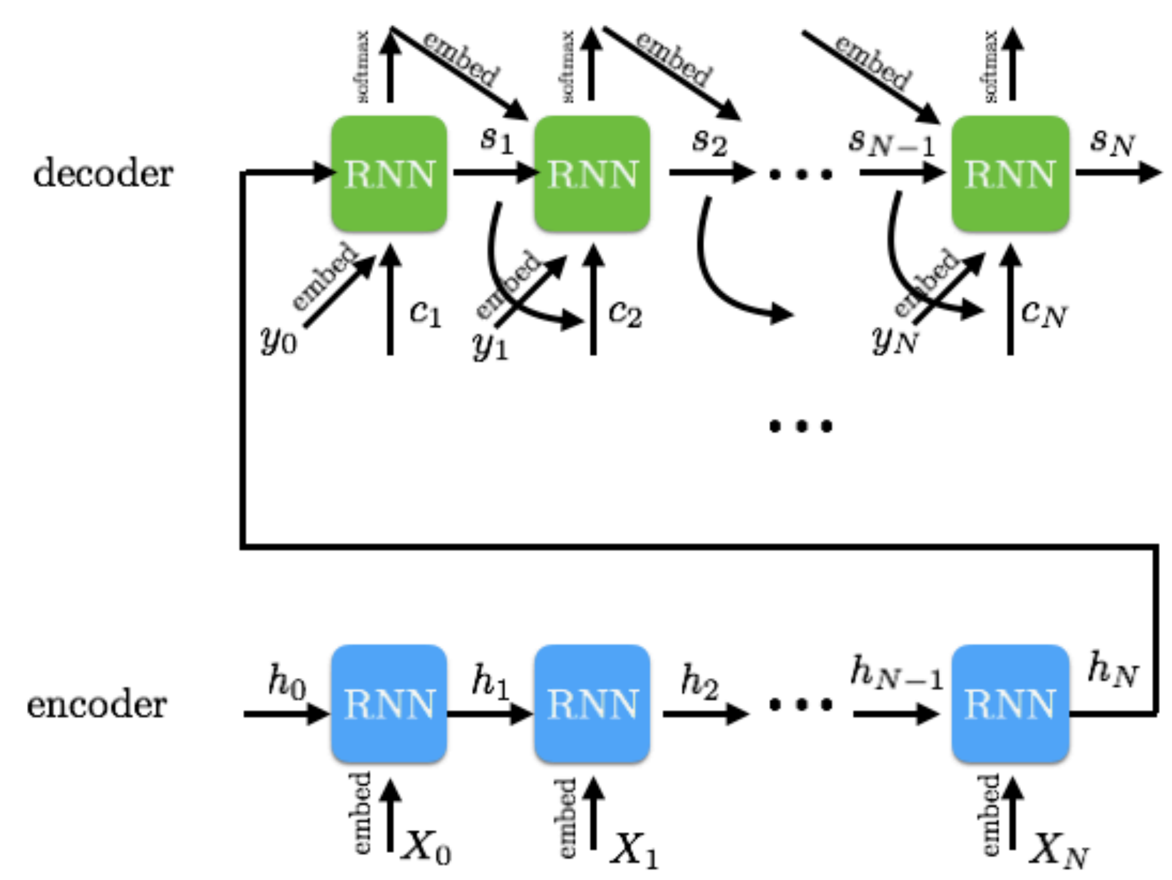
- Each file contains 1024 articles or abstracts
- 183 files in training set
- 61 files in testing and validation sets.





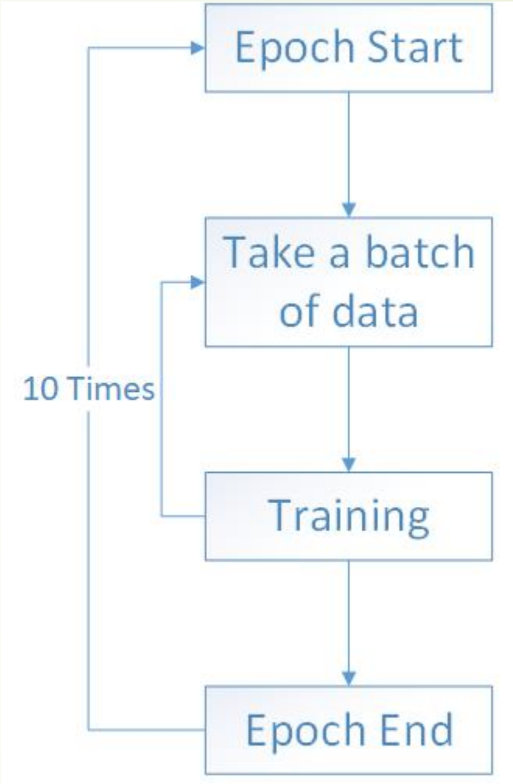
```
57,2,17439,1142,142,49,959,781,1923,5,5489,10,5,16871,26660,4,1640,5,16871,26660,65,188,23
31,5,6569,10,1801,4,35,5,632,26,5,140279,22050,218,55,910,8,19688,1724,4,1036,96,2960,7,8,
23897,332,17,959,30,5,601,31037,22050,4,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
53,2,3430,3293,139,2879,2020,254,20,13,478,10,3654,8,2746,26,8,1370,4,39,20,32,55,48,131,2
043,22,5,4776,6,37,731,107,10,3430,4,20,468,8,247,1361,46,99,534,6,121,6731,11,7674,911,12
,
185,4,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
66,2,1278,385,116,781,514,10,370,1039,22,3170,843,26,18181,4,4102,1065,10631,7335,8471,130
1,14,24,506,580,11,8,152,4,370,74,526,7,249,17,580,1902,22,325,11,283,12,3103,122,3453,4,5
06,64,1449,4879,894,7,8,4664,1382,26,20017,26057,27,35431,56243,19,480,125,4,3,0,0,0,0,0,0
,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
59,2,111747,244,26,5,511,10,5896,30,6060,675,29,619,4,11,7286,6,13,2697,5,6090,1426,2457,1
4,24,5733,7,1836,4,816,2223,164535,11640,6,633,7,4500,106,417,10725,4,111747,25,39,67,32,4
8,306,1959,4,4,35,17287,1124,71,239,528,37,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
53,2,4075,3147,4751,15832,23086,39,26964,7594,6,37,112,9750,3446,38451,1777,7,2778,4,8,200
8,233,8742,2362,2586,53,20,4281,5,836,76,152,6,57,32,48,40680,4,15832,2489,7,24,1214,6,264
,
```

Encoder/Decoder



Training

- 10 epochs
- 4 documents each batch
- Trained 30 hours
- Loss decreases from 12.8 to 6.4



Training

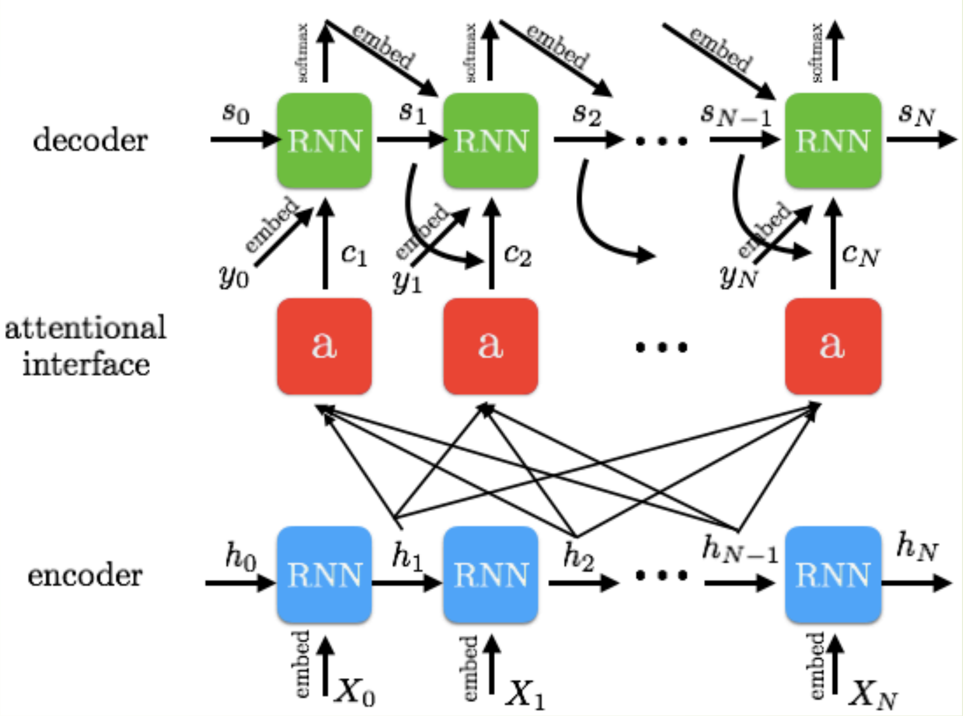
- Improve
  - Data parallel
  - 100 epochs
  - Flexible learning rate

Testing

- Input the article and get predicted abstract
- Evaluate by PyRouge

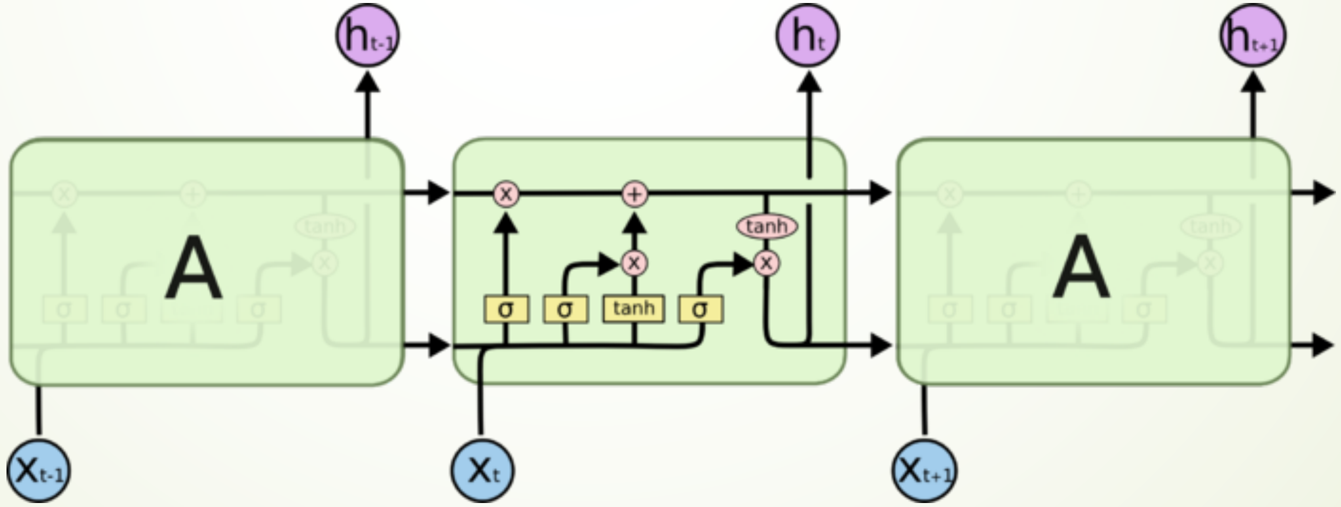
Attention Model

- Based on the attention of human
- Take part of hidden value from encoder



Long-Short Term Memory Model

- Based on human memory mechanism
- Encoder generate multiple hidden values
- Example: “The cloud is in the ___”



Acknowledgements

- Client: Yufeng Ma

References

- Encoder-decoder: https://theneuralperspective.com/2016/11/20/recurrent-neural-networks-rnn-part-3-encoder-decoder/
- Attention model: https://theneuralperspective.com/2016/11/20/recurrent-neural-network-rnn-part-4-attentional-interfaces/
- LSTM model: http://colah.github.io/posts/2015-08-Understanding-LSTMs/