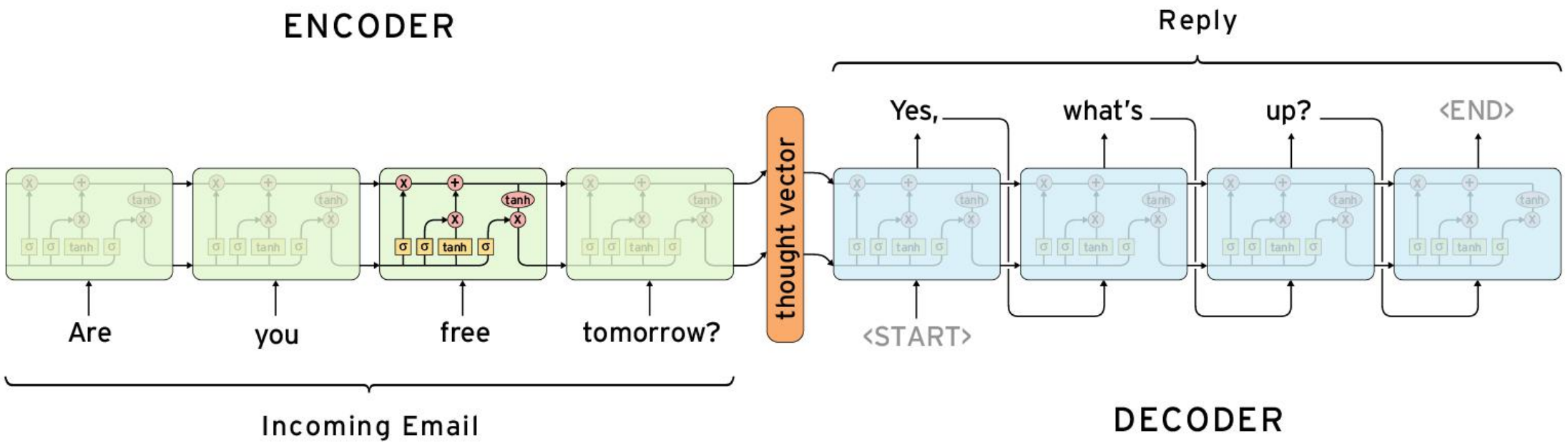
The implementation of Cornell-Movie corpus Chatbot using Seq2Seq

Yiqiang Zhao
June 28 2018



Input Sentence --> Encoder --> Thought Vector --> Decoder --> Output Sentence

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

$Y_1, \ldots, y_{T'}$ = Output Sequence
$x_1, \ldots, x_T$ = Input Sequence
v = Vector Representation

Cornell Movie -- Dialogs Corpus

movie_conversations.txt

u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L194', 'L195', 'L196', 'L197']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L198', 'L199']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L200', 'L201', 'L202', 'L203']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L204', 'L205', 'L206']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L207', 'L208']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L271', 'L272', 'L273', 'L274', 'L275']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L276', 'L277']

movie_conversations.txt

L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!
L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!
L985 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I hope so.
L984 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ She okay?
L925 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Let's go.
L924 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Wow
L872 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Okay -- you're gonna need to learn how to lie.

Data Preparation

Find pairs of conversations

1. change to lower case
2. filter out unnecessary characters
3. filter out too long or too short sentences
4. convert to list of words

1. create words dictionary from sentences
   (Different for pretrained embedding)
2. convert words to indexes

```
q : [You hate me dont you];
a : [I dont really think you warrant that strong an emotion]
```

```
q : [2, 5, 6, 2];
a : [1, 6, 7, 9, 2, 10, 4, 25, 24, 102]
```

Padding

For the variables in the same batch, we need to convert their lengths into fixed length sequences. Here are a few special symbols to fill the sequence.

EOS : End of sentence
PAD : Filler
GO : Start decoding
UNK : Unknown; word not in vocabulary

```
q : ['You', 'hate', 'me', 'dont', 'you', PAD, PAD, ..., PAD]

a : ['I', 'dont', 'really', 'think', 'you', 'warrant', 'that',
'strong', 'an', 'emotion', EOS, PAD, PAD, PAD]
```

Pretrained Embedding vs. Trained Embedding from Scratch

Questions	Answers (Pretrained)	Answers (From Scratch)
you can do it	i ' m not sure .	i dont want to
you are terrible	no , you know .	no im not
i ' m sorry i don ' t know what else to say except im sorry	i don ' t want to talk you .	well thats okay i mean you wouldnt have a very nice time you know what you mean
is something wrong	yes, i ' m sorry .	no no its okay
i m going to miss you	no , i don ' t know .	you wont be here
congratulasions	what ' d you say .	NA ('congratulasions' not in the corpus)
tell me something	no .	you have to go

SHOW ME THE DEMO!

To Improve the Chatbot

- Change the training dataset or incorporate other datasets
- Using bidirectional LSTMs and attention models.
- Tunning hyperparameters
- Try GloVe with more dimensions