Author Identification : A Deep Approach and Comparative Study

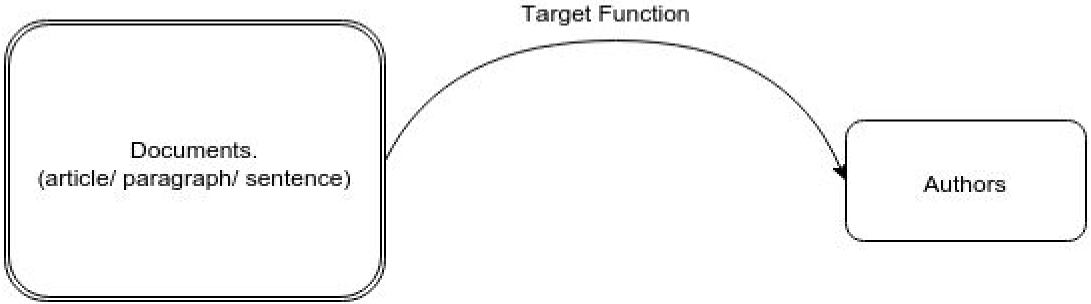
Advisor:
Prof Amitabha Mukherjee

Ankit Pensia (12124)
Anand Pandey (12109)

Problem Statement

Given a text document and a set of authors, learn a function that maps the document to a single author.

Document can be a sentence, paragraph or article.



Previous Work

Earlier work used lexical and grammatical features.

- Bag of Words
- Sentence structure
- Punctuation
- Average Word length

and many more….

Downside:-

- Hand-coded features don’t generalise well.
- Needs a lot of expertise
- Bag of words :- No information about word order is preserved.

DataSet Collected Till Now

Quora (using RSS Feed)

- TopWriters Answers
- 31 Authors and approx 50 answers per author.
- Each answer having 1000 characters.

Will be adding more authors.

Pros :-

Each author has distinct style of writing.

Cons :-

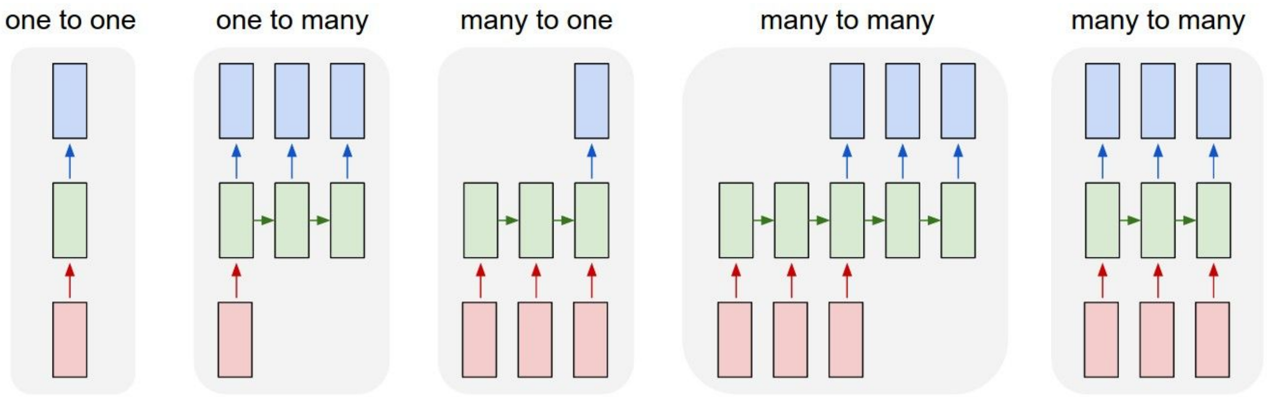
Primary topic of answers vary among authors.

(Technical vs Relationships vs Politics)

******News Article and blogs on specific domain by different authors (Will be collected)**

Recurrent Neural Network

- Mathematics - already presented in other presentations.
- Basically allows to model sequences in any (and many more) of the way below.



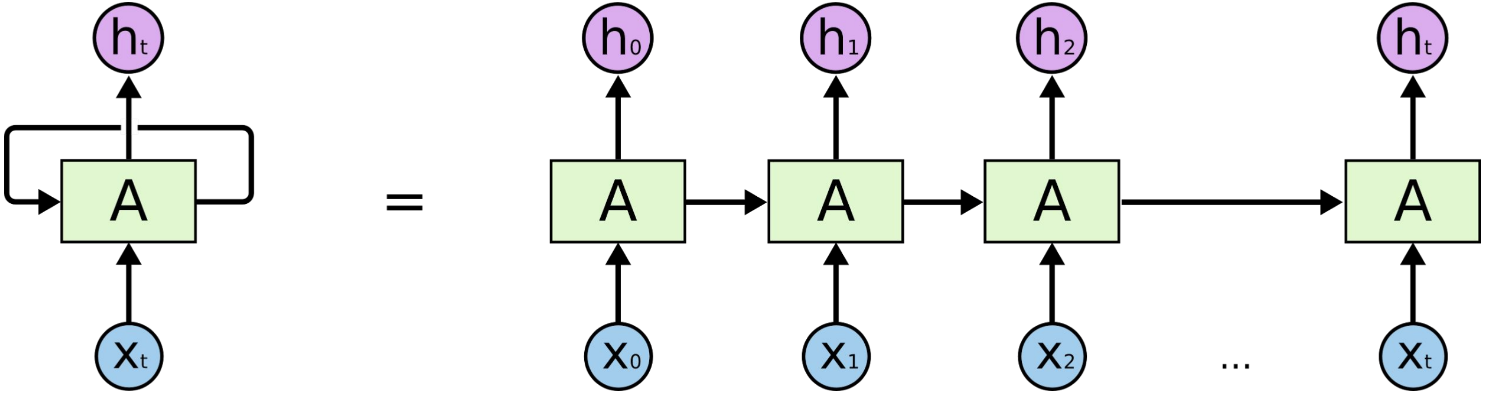
hidden state

input

Image :-http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Recurrent Neural Network.

.



Long Short-Term Memory (LSTM)

- RNNs (traditional architecture) are difficult to train.
  - Neural networks are trained by gradient descents.
  - For RNNs, Gradients either explode or vanish.
-
- If x<1 , gradient doesn’t go back.
- if x>1, gradient explodes.

$$\left\| \frac{\partial h_t}{\partial h_k} \right\| = \left\| \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq (\beta_W \beta_h)^{t-k}$$

LSTM

- One of the variant of RNNs.
- Neuron is replaced by a memory cell.
- Back-propagation works.
- Uses combination of gates.

Image :- https://cs224d.stanford.edu/lectures/CS224d-Lecture7.pdf

LSTM - based Approach.

X0 , X1 ….. Xn are the words and h0,h1...hn are the hidden states of the neurons .

Tree-LSTM

K.S. Tai - et al , 2015

Makes use of the inherent structure, present in the sentences.

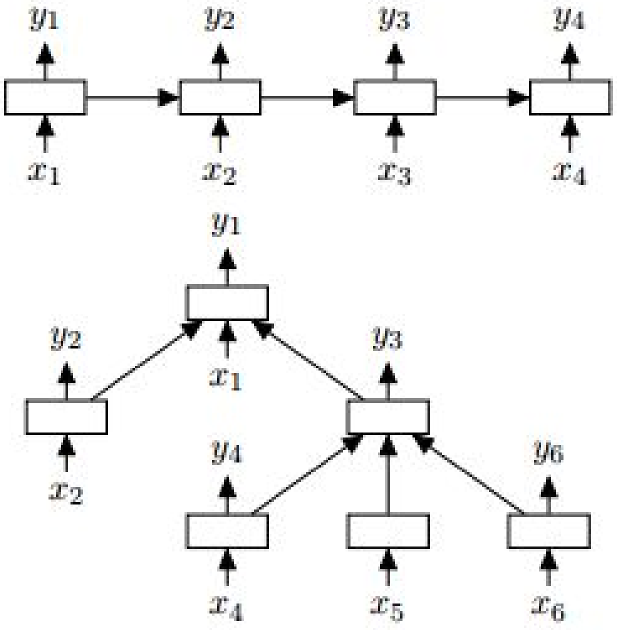
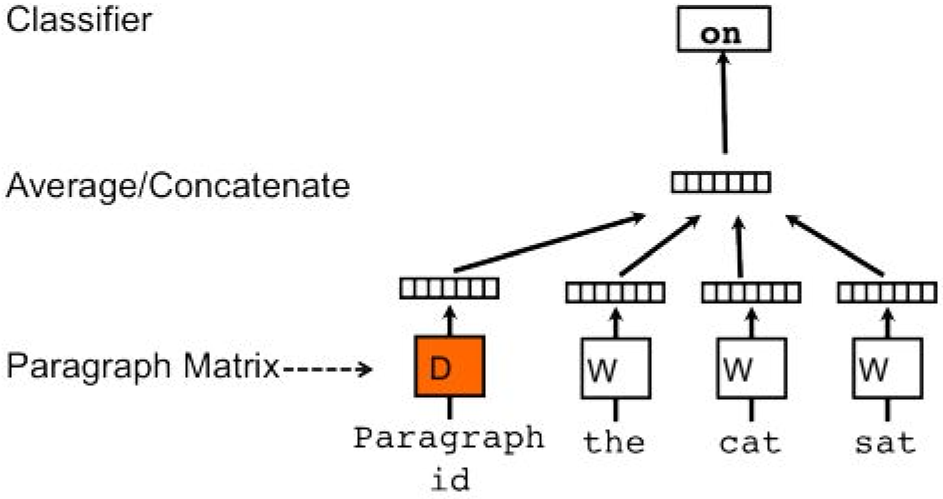


Figure 1: **Top:** A chain-structured LSTM network. **Bottom:** A tree-structured LSTM network with arbitrary branching factor.

Image :- Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks

Paragraph Vectors

- Using authors_id instead of paragraph_id.
- Use similarity metric while inferring.



Distributed Representations of Sentences and Documents (Quoc V. Le - 2014)

Preliminary Results

- Using LSTM - Mean Pooling of hidden layers.
- Just ran a initial version of code on an earlier dataset.
- No fine tuning done (as of now)
- dim_projection(60) and sequence lengths are arbitrarily initialized.
- Whole Answer as a sequence.

Number of Documents in Training (Quora Answers) - 319

Authors - 20

Training and Testing Dataset 70:30.

Dataset	Training	Testing
Top1 Accuracy	0.35	0.12
Top -3 Accuracy	0.72	0.39