Deep Learning Based Authorship Identification

Stanford | ENGINEERING
Electrical Engineering

Chen Qian, Tianchang He, Rao Zhang
(cqian23 / th7 / zhangrao)@stanford.edu

Stanford University

Introduction

Authorship identification has long been a heated issue. It enables us to identify the most likely author of articles, news or messages. Authorship identification can be applied to tasks such as identifying anonymous author, detecting plagiarism or finding ghost writer. Approaches to authorship identification consists of two types: similarity based approaches, such as feature sampling method, and machine learning based approaches, such as SVM. Traditional works perform well on easy tasks, but failed on the scenario where style is not easy to detect, such as news articles.

Problem Statement

The problem consists of two parts: authorship identification and authorship verification. The first part is to identify the author for an article, while the second is to determine whether two given articles are written by the same author.
Deep learning models are resorted to in our work. We built 3 RNN-based model for authorship identification, and a siamese network-based model for authorship verification. Our evaluation methodology is the accuracy and F1 score of the prediction result. Finally we can visualize the result with a confusion matrix.

Datasets and Preprocessing

- Data Sets

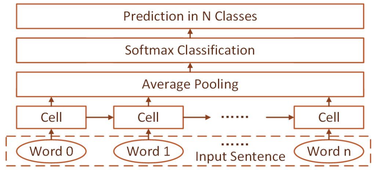
  We used two text datasets for this study: RCV1 C50 and Project Gutenberg. They both contain works of 50 authors. C50 are news articles and Gutenberg contains fictions.

- Pre-processing

  GloVe was adopted as pre-trained word embeddings. All words were trimmed to match the word embeddings. Batch input was enforced to make use of GPU parallel computation. Masks were enabled to eliminate impacts from magic words.

Approach

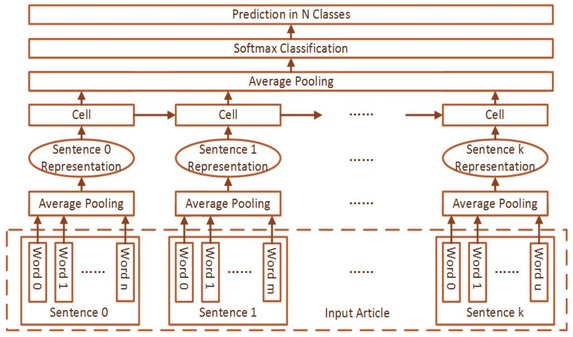
1. Sentence Level GRU



$$\begin{cases} z_t &= \sigma(W^{(z)}x_t + U^{(z)}h_{t-1} + b^{(z)}) \\ r_t &= \sigma(W^{(r)}x_t + U^{(r)}h_{t-1} + b^{(r)}) \\ \tilde{h}_t &= \tanh(r_t \circ U^{(h)}h_{t-1} + W^{(h)}x_t + b^{(h)}) \\ h_t &= (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1} \end{cases}$$

Make authorship prediction based on several sentence. Accuracy could be raised by voting method.

2. Article Level GRU



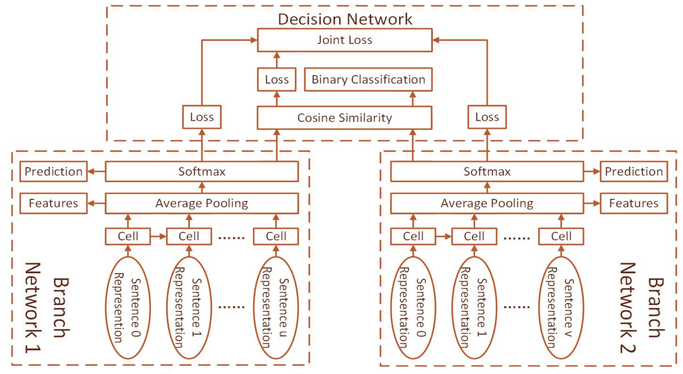
Make authorship prediction based on sentence. Input sentence by sentence, attach equal importance to every word and sentence.

3. Article Level LSTM

$$\begin{cases} i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \\ f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \\ o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \\ \tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1} + b^{(c)}) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\ h_t &= o_t \circ \tanh(c_t) \end{cases}$$

LSTM cell was used as RNN cell.

4. Article Level Siamese Network



$$J = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) - \sum_{j=1}^{n} y_j \log(\hat{y}_j) + \lambda\sigma(wd(f_1, f_2) + b)$$

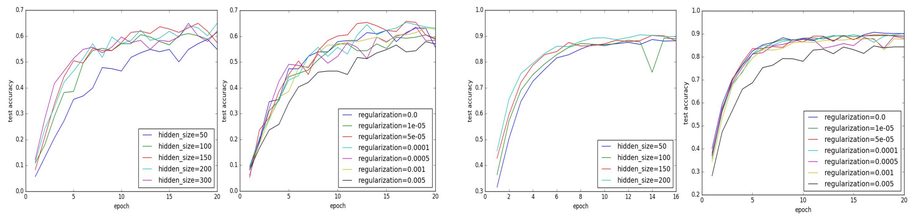
Results

1. Sentence Level GRU

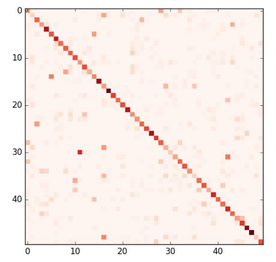
(Not very good, but beyond expectation!)
C50 best accuracy 44%, F1 score 0.43.
Gutenberg best accuracy 53%, F1 score 0.51.
With voting: C50 best accuracy: 63%, Gutenberg: 81%.

2. Article Level GRU

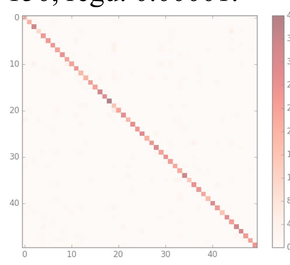
Tested for different hidden size and regularization.



Test accu. for C50. Test accu. for Gutenberg.
> Best combination: hidden size 150, regu. 0.00001.



C50
Best accu: 69.1%
F1 score: 0.66.



Gutenberg
Best accu: 89.2%
F1 score: 0.85

3. Article Level LSTM

C50 best accuracy 62.7%, F1 score 0.604.

4. Article Level Siamese Network

λ	0.0005	**0.001**	0.005	0.01	0.05
Test Accuracy	97.4%	**99.8%**	98.6%	97.4%	96.3%

Future Work

1. Try different size of word embeddings or random initialization.
2. Explore intra-article attention mechanism.
3. Try other models such as MvRNN or CNN.
4. Further analyze the article with the extracted feature.