ConHAN: Contextualized Hierarchical Attention Networks for Authorship Identification

Jean Bouteiller, Victor Jouault, Jack Schooley - **MIT**

I. Introduction

Motivation

Authorship identification is the task of **predicting the author of a given text**. It can be applied to a broad range of task, from ghost writer identification to plagiarism detection.
We take a supervised deep learning approach to this problem, using pre-trained contextualized embeddings and hierarchical attention networks.

Dataset

Dataset: Reuters 50-50 (C50) dataset, a standard benchmark of news article and authors
Size: 50 unique authors, 100 articles per author
Objective: Supervised multi-class classification over 50 authors

Lit. review

We first use [1] as a guide and reproduce their results using GRUs (below are 2 different models):

- Word embeddings initialized with GloVe embeddings
- **Sentence RNN** takes word as input then fed to GRU on words
- **Article RNN** takes averaged sentence-embeddings as input, then fed to GRU on sentences

Target Accuracy: [Qian et al] reach 69% accuracy on C50 dataset

II. New Approach and Models

We leverage **pre-trained contextualized embeddings** via DistilBERT [2] and build several models to create a **document embedding *v*** then used for classification:

A. **CLS DistilBERT**: DBERT's [CLS] token embedding taken as doc embedding
B. **Word Attention DBERT**: Word-attention layer attends to each word separately

We then leverage the article's structure by implementing ideas from the **Hierarchical Attention Networks** framework [3]:

C. **Simple HAN:** Implement 2 layers of attention at word and sentence level
D. **ConHAN**: Add a GRU Sentence Encoder to obtain contextualized sentenced embeddings (figure 2)

III. Results

Methods	Accuracy	F1	Auth Var.
Random Forest	12.6%	8.94%	0.179
RNN [1]	69.1%	46.7%	0.182
CLS DBert	**78.5%**	**78.34%**	**0.173**
Word Att. DBert	77.5%	77.47%	0.178
HAN	75.7%	75.3%	0.187
ConHAN	76.6%	76.36%	0.190

Table 1: Models Results on C50

Pretrained embedding and HAN outperform lit. models but would benefit from more articles

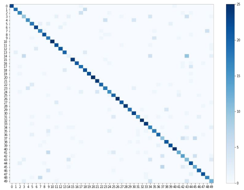

Figure 1: Confusion Matrix for Sentence-RNN HAN

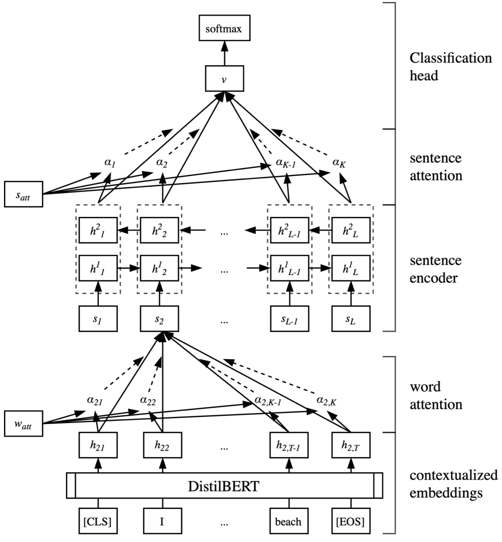

Figure 2: ConHAN architecture

IV. Analysis and Interpretation

III. 1. Model Interpretation - Author Level

We measure if our model is sensitive to authors with rich vocabulary and who are using frequently specific words: correlation between A (author accuracy) and text characteristics

Author i Voc. Richness [4] - K_i
N: Number of distinct words
V(m, N): #words appearing m times

$$K_i = 10^4\left(-\frac{1}{N} + \sum_{m \in N} V(m,N)\left(\frac{m}{N}\right)^2\right)$$

Author i Words Importance - I_i
T_{10}: 10 words with highest *TfIdf*
t(w): Author-level *TfIdf* of word w

$$I_i = \sum_{w \in T_{10}} t(w)$$

$$corr\left(A_{tot}, I_{tot}\right) = -14.5\%$$
$$corr\left(A_{tot}, K_{tot}\right) = -40.8\%$$

→ Our model predicts well authors with poor vocabulary and without important words

III. 2. A Statistical Approach of Model Interpretation

We measure if our model is sensitive to Entities, Part of Speech Tagging, Sentence Length, etc. in articles.
We used a L1-penalized Linear Regression predicting if final model predicts correctly article's label, with elementary features, POS, Entity Recognition, etc.

Feature	Coef	P Value
Intercept	0.756	$< 10^4$
% Conjunction	-0.032	0.013
% Pronoun	0.042	0.005
% Geographical	-0.057	$< 10^4$
% Person	-0.038	0.002
% Noun	0.032	0.028

Figure 3: Regression Coefficients for linear classification

At an article level, our model is agnostic to most features but overfits some patterns

References

[1] Qian et al. 2017. Deep Learning based authorship identification
[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter.CoRR,abs/1910.01108
[2] Yang et al. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics
[3] Tanaka-Ishii, K., & Aihara, S. (2015). Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy. Computational Linguistics, 41(3), 481-502.