WHODUNIT? LEARNING TO CONTRAST FOR AUTHORSHIP ATTRIBUTION

Bo Ai[1] [bo.ai@u.nus.edu], **Yuchen Wang**[1] [yuchen_wang@u.nus.edu], **Yugin Tan**[1] [tan.yugin@u.nus.edu], **Samson Tan**[2] [samson@amazon.com]
[1]**National University of Singapore**, [2]**AWS AI Research & Education**

Motivation

Authorship Attribution (AA) is the task of identifying the author of a given text.
Why is it important?

- Various downstream applications: anonymous email author identification, plagiarism detection, forensic investigation, machine-generated text detection
- *Turing Test* for Natural Language Generation models

The key is finding representations that can differentiate between authors.

Existing methods achieve this via feature engineering:

- Style-based features: letter frequency, punctuations, word length
- Content-based features: Word N-grams
- Hybrid features: Character N-grams

What are the drawbacks?

- Dataset-dependent engineering, performance inconsistent across corpora
- Performance does not scale well with data

Approach Overview

We adopt a data-driven approach to learn highly author-specific representations. Specifically, we propose two strategies:

- Instead of learning a model from scratch, we exploit the general representations in pretrained Large Language Models (LLMs)
- We use a contrastive learning objective to finetune the model to capture the idiosyncrasies of each author

We evaluate our approach on

- Three AA benchmark datasets: Blog10, Blog50, IMDb62
- One recently proposed corpus containing texts from both machine and human writers: TuringBench

Our contributions:

- The first attempt to integrate contrastive learning with pre-trained language model finetuning for AA
- Advanced the state-of-the-art across datasets
- An analysis of the learned representations to study its strengths and reveal potential ethical concerns

Methodology

Setup

- The classification model p, factorized as $\mathrm{p} = \phi \circ h$, where ϕ is a feature extractor
- The dataset D, a sample is a text-author pair $\langle t, a \rangle \in D$

The objective is to maximize predictive accuracy

$$Acc = \mathbb{E}_{\langle t,a \rangle \in D} \, \mathbb{1}_{argmax(p(t))=a}$$

The traditional approach: cross-entropy loss

$$\mathcal{L}_{CE} = -\sum_i a_i \log(p(t)_{a_i})$$

Our idea: construct a representation space in which

- Similarity between texts from the ***same*** authors is ***maximized***
- Similarity between texts from ***different*** authors is ***minimized***

We propose a contrastive learning objective

$$\mathcal{L}_{CL} = -\sum_i \log\left(\frac{\sum_{a_i=a_j} \exp(\mathbb{S}_{i,j}/\tau)}{\sum_k exp(\mathbb{S}_{i,k}/\tau)}\right)$$

where $\mathbb{S}_{i,j}$ is a similarity matrix that satisfies

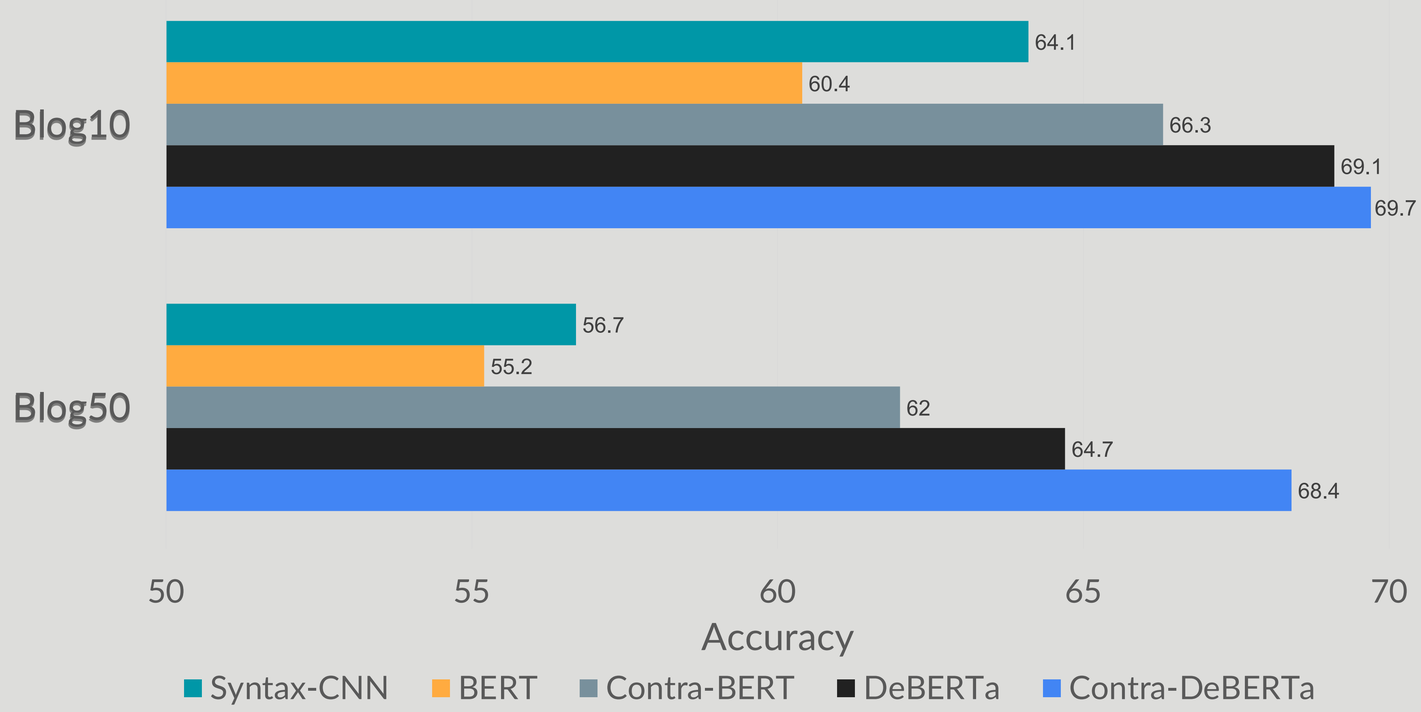
$$\mathbb{S}_{i,j} = \cos(e_i, e_j) = \frac{e_i * e_j}{\|e_i\|\|e_j\|}$$

A pre-trained language model is finetuned to optimize the joint loss

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda * \mathcal{L}_{CL}$$

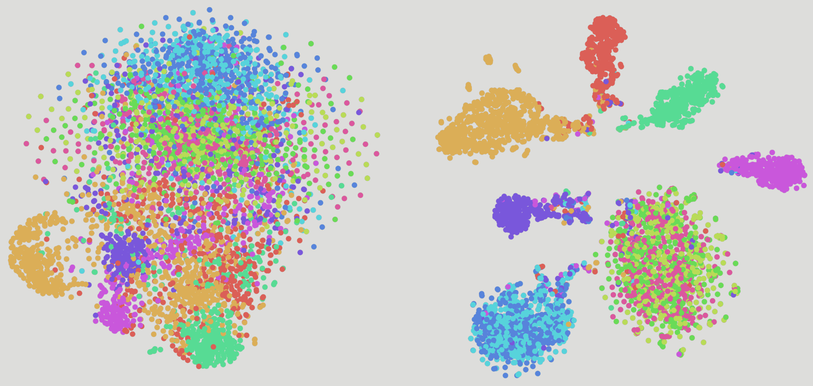
Results

Here we present results on human-writer datasets Blog.
For more results, please refer to the paper.



Discussions

- Our contrastive objective consistently improves upon vanilla finetuning across different data regimes. It is robust to scare data while scalable to huge corpus.
- The propose objective indeed creates a more compact but distinct representation for each author.



Visualization of features from BERT (left) and Contra-BERT (right) on Blog10 dataset

- With the objective, the model may learn to increase the accuracy on some authors at the cost of sacrificing the performance on other authors – addressing this would not only improve model performance but alleviate ethical concerns