Learning Multi-Domain and Cross-Domain Authorship Representation

Yang Su
Cornell
ys724@cornell.edu

Sasha's NLP Group Members *
Cornell
placeholder@cornell.edu

Abstract

Authorship identification and attribution aim to identify the belongings of the given text from a set of known authors. Previous approaches tried to learn author-level embeddings via contrastive learning that can be transferred to multiple domains that the author has written content about, but failed to give satisfactory results. We first scale the contrastive learning batch size beyond GPU memory constraint by using more negatives in each training batch and a larger pre-trained model backbone, then propose a data sampling and augmentation technique that greatly improves previous state-of-the-art results on multiple large-scale datasets, incorporating hard-positive and negative examples during in-batch sampling, and further augmenting this data by fine-tuning a generation model that produces the missing hard text corpora. We find that this method enables the model to focus its attention less on topic-related tokens of the authors, and more on the combination of punctuation and semantic properties, which is where its main performance improvement comes from.

1 Introduction

Terabytes of content are produced by anonymous users every day, which inherently contain individual portrait information. This poses a risk as malicious agents exploit this data to uncover user identities or fabricate content falsely attributed to the authors, thereby influencing public opinion. We are hence motivated to protect authorship privacy by developing autonomous systems to detect the authorship of the given content. Our study is confined to digital text, which is also known as authorship attribution Tyo et al. (2022). This task has practical applications in literary studies, digital forensics, and intellectual property rights, and is crucial for detecting social media accounts and combating online disinformation campaigns. Our work is served as one of the performers in the HIATUS (Human Interpretable Attribution of Text Using Underlying Structure) competition, where teams compete to generate higher fidelity representations between individual authors' unique linguistic fingerprints.

Prior research in authorship attribution has mainly focused on deriving author-level and document-level embeddings through contrastive learning objectives Soto et al. (2021). However, these methods often falter with texts of varied genres or works by authors with diverse themes, resulting in less-than-ideal outcomes. To overcome these challenges, our paper presents an innovative approach that significantly enhances performance across multiple large-scale datasets[1]. Our key contributions include:

- We introduce a hard example sampling strategy to incorporate hard-positive and hard-negative examples during in-batch sampling, and then further augment this data by fine-tuning a generation model to produce the missing hard text corpora, with carefully designed supervised fine-tuning objectives.

*Corresponding Authors.

[1] Link to our project presentation: Google Slides.

- We additionally scale up the contrastive learning batch size to surpass GPU memory constraints, allowing for the inclusion of a substantially larger number of negative samples in each training batch and a much larger pre-trained model backbone.
- We produce new state-of-the-art results, especially on cross-domain datasets where both the genres and the authors' writing styles are unseen. The model also demonstrates superior performance with longer text contexts, indicating its robustness in handling more extensive text corpora.

2 Related Work

Authorship Attribution Also referred to as authorship identification and closely related to authorship verification, the task aims to determine whether two documents come from the same author. The input of the problem would be a large group of authors with their written text corpus, and the output would be author-level and document-level embeddings. The previous state-of-the-art approach to this task Soto et al. (2021) applied a contrastive learning objective to match texts that come from the same author, which aims to derive feature space that captures author-level variation across diverse text content written in different genres. It achieves satisfactory performance when the genre of the dataset is unique, but not when the genre is mixed from the same author or there are multiple genres produced by a group of authors.

Memory-Efficient Contrastive Training Contrastive training quality has been shown to depend on its effective batch size, but training with a sufficiently large batch size is generally not feasible with limited GPU memory. The method of *gradient caching* Gao et al. (2021), a special use of *gradient checkpointing* in contrastive training, can scale the batch size of contrastive form loss and increase the in-batch negative examples far beyond GPU RAM limitation, with little slow down in the training time of the model. Intuitively, it separates the backpropagation from contrastive loss to the output representation part and the representation to the encoder model part. An extra forward pass without gradient tracking is performed to compute the representation and store its gradient in the cache, which removes the gradient dependency in the backpropagation of the encoder, and we can then update the gradient of the encoder one sub-batch at a time to fit in limited GPU memory.

3 Methods

Model Our model architecture follows the one proposed in the LUAR paper Soto et al. (2021), where we learn a function mapping that takes in a minibatch of N different authors, each writes $2 * M$ documents, and the output is a pair of author-level embeddings (called *queries* and *targets*, each with N corresponding authors) that come from M non-overlapping documents of each author. The model is then fitted with the contrastive learning objective, where two embeddings are close if they come from the same author, and vice versa, optimized with the InfoNCE loss.

Hard In-Batch Sampling Previous approaches work well on single-domain datasets where the contents are within the same text genres but do much worse when (1) the texts written by different authors contain a high level of topic diversity, or (2) the authors themselves write different topics (more generally, they perform different *actions*) within their written texts. With feature importance analysis of the models' results (to do: add explainability section), we find that when the model focuses too much on topic-related tokens in the text, it becomes hard to separate texts with similar contents from different authors, and also challenging to distinguish texts written by the same authors, but contains different contents. To accommodate this issue, we implement a *single-domain-in-batch* strategy where each batch consists of documents from a single domain. This approach introduces harder negatives into the training process, i.e. the authors to distinguish from all write similar contents in a certain domain, making the training procedure more challenging. Furthermore, we incorporate a *single-action-in-batch* strategy, which groups documents from a single action of each author together. This strategy introduces both harder negatives and positives, i.e. the texts from the same authors contain contents from different domains, additionally to previous hard sampling procedures.

Hard Example Fine-Tuning and Generation In the previous approach, we can find that restricting all queries and candidates to come from the same action, while ensuring the largest action-wise difference between queries and candidates could theoretically yield the hardest training batch. However, this requires a batch of authors to have at least 2 genres in common, and these 2 genres have to be far away from each other, which is very unlikely to be found in the datasets, especially as we will scale up the training batch size later on. Hence, we propose a simple yet effective data augmentation technique to generate the hard negatives with the missing genres. The detail of the process is shown in Figure 1. We apply instruction fine-tuning to a generation model (usually a decoder model with the next-token-prediction objective), while only back-propagating the loss on its output (the blue part). This method enables us to fill in any genre we want to use during training by creating synthetic data.

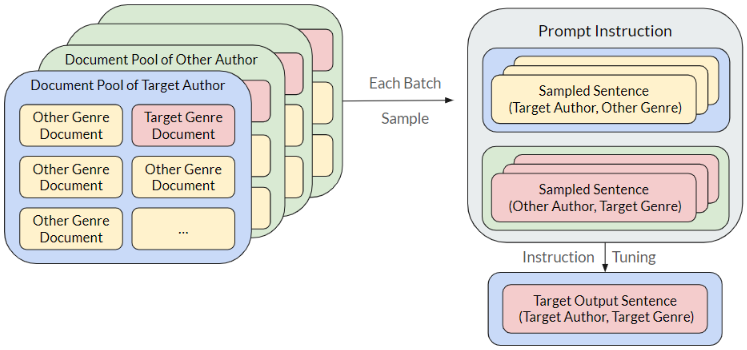
Document Pool of Other Author
Document Pool of Target Author
Other Genre Document
Target Genre Document
Other Genre Document
Other Genre Document
Other Genre Document
...
Each Batch
Sample
Prompt Instruction
Sampled Sentence (Target Author, Other Genre)
Sampled Sentence (Other Author, Target Genre)
Instruction
Tuning
Target Output Sentence (Target Author, Target Genre)

Figure 1: The overview of Hard Example Fine-Tuning and Generation. During each training batch, we sample a target (author, genre) pair and treat that as the target output, then sample sentences from pairs of (matched author, unmatched genre) and (unmatched author, matched genre) and apply proper instructions to fine-tune the model.

Scaled In-Batch Sampling While boosting the model's performance with a much larger batch size might be helpful, another side effect is that the peak memory the model requires now only depends on the sub-batch size we use, which can be reduced to 1. This is important as we can now train the model with the largest pre-trained backbone that can fit into the GPU. As we will see in Section 6, the improvements from the previous two strategies are more evident with the help of scaling up, compared to the same scaling technique applied to the baseline model.

4 Datasets

We conduct experiments on datasets from a variety types of domains and topics, including reviews, posts, blogs, and books. For each dataset, a random portion (20%) of disjoint authors are split to form the evaluation set. The detailed dataset information is in Table 3.

Reddit Comments This is a filtered version of the PushShift Reddit Dataset Baumgartner et al. (2020) with 1.2 million authors, each has written at least 100 posts (documents). It has the shortest token length per document on average and the most number of authors and documents to distinguish from, hence is also the largest dataset we use that compromises over 95% of the total training corpus. It also contains the most number of genres (125k), categorized by subreddits, where each author posts in 28 subreddits on average.

Fanfiction Stories A filtered version of the PAN 2020 Dataset Wiegmann et al. (2019) with over 270k long-form stories written by 40k authors. There are 1.6k story types, but each author writes only 1.24 types of stories on average.

Single-Genre Datasets Amazon Reviews is a review collection on Amazon products Ni et al. (2019) with 130k authors, along with other 5 considerably small datasets (with less than 10k authors in total), they all have no metadata, thus no cross-genre information that can be directly extracted from the documents.

We perform Personal Identifiable Information (PII) removal on all datasets (except fanfiction) because we noticed that the model will likely focus on entity names in its training and cause overfitting issues.

5 Experiment Setup

Metrics We evaluate performance with two popular retrieval metrics. *Rank @k*: given a query and group of target author embeddings, the probability that the matched target author embedding is within the top-k closest embeddings to the query. *MRR (Mean Reciprocal Rank)*: the average reciprocal ranks across all queries and first matched targets.

6 Results

Hard In-Batch Sampling See Table 2. Our evaluation has one difference from the original LUAR paper, where we ensure determinism by always taking the first 32 tokens of each document from each author. Firstly, notice that the original LUAR approach trained on multi-domains (8-Domain) is giving worse results when evaluated on multi-domains compared to solely trained on the Reddit dataset, although they did perform better on cross-domain datasets (amazon and fanfiction). In contrast, our approach with *single-domain-in-batch* gives better performance in all datasets, both in and cross-domain, suggesting the generalization ability of our approach.

Scaled In-Batch Sampling We successfully make *Gradient Caching* compatible with PyTorch Lightning, the original framework used by LUAR to enforce the reproduction of their results and see performance improvements from the training loss. However, the scaled training process is very long (expected more than 1 month for 1B+ parameter model backbone) with our current hardware (8 A6000s). Thus, We will continue to run and meanwhile look for better methods to scale up our methods.

Our cross-domain performance is shown in Table 1, where the evaluation datasets (blind, from the HIATUS project team) consist of 5 held-out domains and a mixed-domain dataset, containing both unseen genres and authors' writing styles.

Model	Metric	Evaluation Datasets						Average
		HRS 1.1	HRS 1.2	HRS 1.3	HRS 1.4	HRS 1.5	Cross Genre	
LUAR	MRR	55.6	25.4	15.0	29.4	28.1	5.3	26.5
	R@8	67.1	29.3	16.6	31.9	27.5	2.3	29.1
Ours	MRR	**72.4**	**53.5**	**34.7**	**33.0**	**38.9**	**12.1**	**40.8**
	R@8	**84.0**	**63.7**	**39.1**	**36.4**	**39.1**	**7.9**	**45.0**

Table 1: Evaluation results from our complete model vs. the LUAR model.

7 Future Steps

We plan to do an explainability analysis to test our hypothesis as to whether our model indeed performs better by focusing more on non-topic but stylometric patterns of the authors, while investigating the effect of dataset topic diversity on the performance. We will also try other hard example mining strategies including retrieval-based sampling, sentence-level hard sampling, etc. There is also potential to combine the hard example generation model and the retrieval model by iteratively training them together via an RLHF-like fashion, which we will leave as future work.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, 2021.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Rafael A. Rivera Soto, Olivia Miano, Juanita Ordonez, Barry Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In *EMNLP*, 2021.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*, 2022.

Matti Wiegmann, Benno Stein, and Martin Potthast. Overview of the celebrity profiling task at pan 2020. In *Conference and Labs of the Evaluation Forum*, 2019. URL `https://api.semanticscholar.org/CorpusID:198488981`.

A Appendix

			Training Datasets					
			Reddit		Fanfic		8-Domain [2]	
			R@8	MRR	R@8	MRR	R@8	MRR
Evaluation Datasets	LUAR (Original)	Reddit	65.61	50.37	12.10	7.64	/	/
		Amazon	68.91	55.59	28.91	20.06	/	/
		Fanfic	41.58	30.61	50.89	41.20	/	/
	LUAR (Reproduce)	Reddit	77.40	62.84	15.17	9.92	73.32	57.93
		Amazon	69.10	55.97	31.79	22.63	74.41	60.31
		Fanfic	8.06	5.03	9.17	5.57	15.17	9.24
		8-Domain	69.29	56.13	17.68	11.83	67.85	53.60
	Single-Domain-In-Batch (Ours)	Reddit	/	/	/	/	76.77	62.24
		Amazon	/	/	/	/	82.41	70.02
		Fanfic	/	/	/	/	13.01	7.74
		8-Domain	/	/	/	/	71.53	58.32
	Single-Action-In-Batch (Ours)	Reddit	**77.49**	**63.10**	**17.63**	**11.49**	**76.85**	**62.36**
		Amazon	**71.04**	**58.28**	**44.20**	**31.86**	**82.93**	**70.76**
		Fanfic	**9.03**	**5.56**	**19.89**	**13.31**	**22.94**	**15.39**
		8-Domain	**69.79**	**56.77**	**22.66**	**15.35**	**72.67**	**59.29**

Table 2: Results from hard in-batch sampling on different training and evaluation datasets.

Dataset	A	D	D/A	T/D	X_{pii}	X_{genre}	G	G/A	A/G	D/G
Reddit Comments	1.07M	312M	300	46	✓	✓	125k	28.4	243	2.57K
Amazon Reviews	108K	19.3M	179	113	✓		/	/	/	/
Fanfiction Stories	262K	31.7M	121	76		✓	1600	1.24	203	12.5K
Blog Authorship	1100	304K	276	264	✓		/	/	/	/
Enron Emails	86	90.1K	1048	98	✓		/	/	/	/
IMDB1M Reviews	34	10.6K	311	345	✓		/	/	/	/
NYT Comments	2244	557K	248	107	✓		/	/	/	/
Yelp Reviews	3492	695K	199	193	✓		/	/	/	/

Table 3: An overview of the datasets (training split). A is the number of authors, D is the number of documents, D/A is the average number of documents for each author, T/D is the average number of tokens for each document. X_{pii} denotes whether the dataset is preprocessed by Personal Information Removal (PII), X_{genre} denotes whether the dataset has cross-genre metadata that allows us to do *single-action-in-batch* sampling. G is the number of genres, G/A is the average number of genres written by each author, A/G is the average number of authors that have written documents in each genre, D/G is the average number of documents with each type of genre.