Improving Efficiency and Robustness of Transformer-based Information Retrieval Systems

Edmon Begoli
begolie@ornl.gov
Oak Ridge National Laboratory (ORNL)
Oak Ridge, Tennessee, USA

Sudarshan Srinivas
srinivasans@ornl.gov
Oak Ridge National Laboratory (ORNL)
Oak Ridge, Tennessee, USA

Maria Mahbub
mahbubm@ornl.gov
Oak Ridge National Laboratory (ORNL)
Oak Ridge, Tennessee, USA

ABSTRACT

This tutorial focuses on both theoretical and practical aspects of improving the efficiency and robustness of transformer-based approaches, so that these can be effectively used in practical, high-scale and high-volume information retrieval (IR) scenarios. The tutorial is inspired and informed by our work and experience while working with massive narrative datasets (8.5 billion medical notes), and by our basic research and academic experience with transformer-based IR tasks. Additionally, the tutorial focuses on techniques for making transformer-based IR robust against adversarial (AI) exploitation. This is a recent concern in the IR domain that we needed to take into concern, and we want to want to share some of the lessons learned and applicable principles with our audience. Finally, an important, if not critical, element of this tutorial is its focus on didacticism – delivering a tutorial content in a clear, intuitive, *plainspeak* fashion. Transformers are a challenging subject, and, through our teaching experience, we observed a great value and a great need to explain all relevant aspects of this architecture and related principles in the most straightforward, precise and intuitive manner. That is the defining style of our proposed tutorial.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

transformer neural networks, information retrieval, efficient transformers, robust transformers

ACM Reference Format:
Edmon Begoli, Sudarshan Srinivas, and Maria Mahbub. 2022. Improving Efficiency and Robustness of Transformer-based Information Retrieval Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3477495.3532681

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532681

1 MOTIVATION

This tutorial is based on our experience with developing information retrieval (IR) systems over some of the largest medical narrative datasets (VA's 8.5 billion medical notes dataset [4]), and a need to have reliable, efficient and well-performing infrastructure to conduct intra-disciplinary, large-scale medical research related to critical medical problems such as prostate, lung and liver cancers; suicide prevention; and cardio-vascular diseases. To develop and support such system, we had to evolve several versions of the IR pipeline and re-train and fine-tune the transformer/BERT-based models hundreds of times. With the stated scale of medical notes and the need for frequent re-training and fine-tuning, we had to devise efficient methods for training and fine-tuning as well as for the most efficient use available of heterogeneous resources (e.g., multiple versions of NVIDIA GPUs and CUDA releases) in order to have an IR pipeline that can effectively support our program.

These experiences and the lessons learned motivated us to put this tutorial together. Furthermore, we teach a graduate-level "Special Topics in NLP" class and we conduct research where we examine problems such as efficiency and robustness of the state-of-the-art NLP models. All this motivated us to put together a tutorial that would inform the beginner-to-intermediate IR community on the methods and techniques for improvements relevant to IR and based on the transformer networks.

2 OBJECTIVES

Transformers are a powerful neural architecture that has advanced the state-of-the-art to near or superhuman performance on many of the tasks relevant to information retrieval (auto-summarization, search, reading comprehension, machine translation, etc.)

However, transformer-based architectures are computationally complex and can be resource-consuming to train and fine-tune, especially for long-context NLP tasks. Recently, new approaches have been developed aimed at improving the efficiency of transformers' training and fine-tuning, especially at making self-attention and other critical components run more efficiently.

Specifically, the tutorial will focus on techniques and principles for efficient computation and efficient, light, and fast transformers - model size reduction and working with efficient self-attention.

The tutorial will describe how to make efficient transformer-based IR systems using distillation, pruning, and quantization. We will review efficient sparse transformers such as Linformer [8], BigBird [9], Reformer [3] and Performer [1], and we will review the impacts on the critical system performance metrics such as memory consumption, model size and runtime performance.

For model size reduction, we will cover the following approaches:

2.1 Knowledge distillation

Knowledge distillation is a process of transferring "knowledge" from a bigger to a smaller model. This process conceptualizes *teacher* and a *student* model, where the knowledge from a larger *teacher* model is transferred to a smaller, *student*, model [6]. For the tutorial, we cover both fine-tuning knowledge distillation techniques (augmenting ground truth labels from the teacher to student to learn from) and pre-training ones (e.g. transference of the knowledge about transformers' masked language models from teacher to the student).

2.2 Weight Pruning

Weight pruning is a model compression techniques that removes the weights from the model that contribute nothing or very little to model's results. In the context of IR, weight pruning is useful if we want to deploy a model with a lesser computational complexity and intensity of the computation on, for example, smaller platforms such as mobile devices, assistive technologies, etc. For the tutorial, we will review some of the best established methods (use of score matrix, magnitude (unstructured) pruning, weight pruning, etc.), and the use of related optimization libraries (e.g., Neural Networks Block Movement Pruning [5]). We discuss the implications of pruning for IR-relevant tasks, consequent performance, and the effects of pruning on different GPU architectures.

2.3 Quantization

Quantization [2] is an optimization technique that aims to reduce the numerical representational precision where that is feasible. In the context of transformer networks, the technique aims to identify and reduce the representational size of the weights in the network. For example, if we have a network with parameters that are stored in in 16-bit variables but the informative bits are all stored in the first 6-8 bits, we will reduce the representation of the weight/parameter to a 8-bit variable, hence saving 50% space and consequent computational complexity. For the tutorial, we will present three kinds of *quantization* (dynamic, static and quantization-aware training), illustrate the approaches with PyTorch *QFunctional* wrapper class, and present the resulting benchmarks – all this with supporting code.

2.4 Robustness

Finally, we recognize that language models can also be subjects to adversarial exploitation and we will present some fundamental techniques for making IR-oriented transformer-based pipelines more robust with respect to adversarial exploitation (e.g., data sanitation defenses [7]).

2.5 Didacticism

Finally, an important, if not critical, element of this tutorial is its focus on didacticism – a focus on delivering the content in a clear, intuitive, plainspeak fashion. Our experience in teaching graduate courses as well as working in applied setting informs us that both research and applied community could benefit from clearly explained transformer-related topics, including transformers themselves and we plan to deliver this tutorial in that fashion.

3 RELEVANCE TO THE INFORMATION RETRIEVAL COMMUNITY

While not answer to all IR research and application problems, transformer-based models (variants of BERT, other models) have been effectively used in practice, while achieving the state-of-the-art performance in search, neural machine translation, auto-summarizing, and reading comprehension/question answering research and applications. Consequently, knowing how to efficiently use transformers for these tasks, especially when working with very large corpus of narratives, is a common IR challenge and this tutorial intends to focus on those kinds of problems (practical, applied, large corpus). Further, to fully appreciate the proposed approaches one needs to understand the transformer architecture. For that reason, the first sections of the tutorial give an intuitive and complete overview of the transformer architecture which should be of great benefit for the IR audience not yet intimately familiar with the inner workings of transformers and their essential mechanisms (which are in themselves of relevance to IR research).

4 RELATED PAST TUTORIALS

This tutorial is a distillation of the practical experience and the consequent lessons learned working on the project between the US Department of Veterans Affairs (VA) and The Department of Energy (DOE) called MVP CHAMPION. As part of that program, we have offered number of similar tutorials to academic and industrial participants on the programs. Also, we base some the content of this tutorial on the content and experience in teaching upper-graduate classes that Dr. Begoli teaches at the The University of Tennessee, Knoxville – (*Special Topics in Natural Language Processing (with Transformers)*, and *Adversarial Learning)*. In terms of other related tutorials offered by other researchers and other conferences, we are not aware of any tutorial that focuses on this specific subject. In the past years, SIGIR and other conferences have featured tutorials on transformers for IR. There was a significant number of research papers on the subjects of efficient transformers, but we are not aware of any tutorial that covers both efficiency and robustness in the context of IR and transformers and that presents it at the beginner-to-intermediate audience level.

5 FORMAT

The tutorial will be delivered as an in-person presentation combining review of the concepts as well as going over the practical code examples. Specifically, the presentation will cover the overview of the conceptual foundations (overview of the transformers and transformer architectures, efficiency and robustness algorithms); review of specific optimization techniques by discussing the algorithms and going over the specific code examples. This will be followed by 30 minutes of live questions and answers (Q&A) session. Two co-authors will be remote and they will facilitate the remote questions and answers.

6 DETAILED SCHEDULE

Figure 1 presents a detailed schedule for the tutorial. The schedule follows the intended structure for the tutorial where we want to start with a general background on the transformer architectures and the relevant IR application in the first hour/hour and a half,

and then spend second hour and a half presenting and discussing specific efficiency optimization and robustness techniques. We close the tutorial with thirty minutes questions and answer session.

Content	Duration	Description
Introduction to transformers	30 min	Introduce transformer architecture, key mechanisms, and components.
Applications in IR	30 min	Applications in neural machine translation, auto-summarization, reading comprehension/ question answering, etc.
Break	10 min	
Efficiency Improvement Principles and Techniques	60 min	Discussion and overview of the knowledge distillation, pruning and quantization techniques.
Break	10 min	
Improving Robustness	30 min	Discussion and overview of the techniques that improve the robustness of transformer models against adversarial (AI) exploitations.
Q&A	30 min	Question and answering session with participants.

Table 1: Tutorial schedule and content.

7 SUPPORTING MATERIALS

The tutorial will be delivered live, by going over slides and python code in Google Colab **notebooks**. All of this material will be available to the attendees as freely downloadable **PDF of the presentation** and the shared Colab notebook with code samples. Material will be available in Begoli's Github repository[1].

ACKNOWLEDGMENTS

This manuscript has been in part co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy, and under a joint programs (MVP CHAMPION and VICTOR), between the U.S. Department of Energy (DOE), and the U.S. Department of Veterans Affairs (VA). Part of this research (academic) was supported by Google Research resources made available to Dr. Edmon Begoli.

[1] https://github.com/ebegoli/SIGIR2020-Efficient-Transfomers

REFERENCES

[1] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794* (2020).

[2] Tianchu Ji, Shraddhan Jain, Michael Ferdman, Peter Milder, H Andrew Schwartz, and Niranjan Balasubramanian. 2021. On the Distribution, Sparsity, and Inference-time Quantization of Attention Values in Transformers. *arXiv preprint arXiv:2106.01335* (2021).

[3] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* (2020).

[4] Kathryn E Knight, Jacqueline Honerlaw, Ioana Danciu, Franciel Linares, Yuk-Lam Ho, David R Gagnon, Everett Rush, J Michael Gaziano, Edmon Begoli, Kelly Cho, et al. 2020. Standardized architecture for a mega-biobank phenomic library: the million veteran program (MVP). *AMIA Summits on Translational Science Proceedings* 2020 (2020), 326.

[5] François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. *arXiv preprint arXiv:2109.04838* (2021).

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[7] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 2017. Certified defenses for data poisoning attacks. *Advances in neural information processing systems* 30 (2017).

[8] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* (2020).

[9] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems* 33 (2020), 17283–17297.