7164CEM

Individual Research Project Preparation

2324JANMAY

Project Proposal and Initial Literature Review

Sujan Tumbaraguddi

Student ID: 14194733

Table of Contents

1. Project Title

Knowledge Distillation from Generative Language Models

2. Project Topic

Investigate knowledge distillation techniques to transfer knowledge from large generative language models (e.g., GPT-3, BERT) to smaller, more efficient task-specific models for natural language processing (NLP) tasks.

3. Research Objective

To investigate methods for knowledge distillation that can be used to move knowledge for natural language processing (NLP) tasks from larger, more complex generative language models (like GPT-3 and BERT) to smaller, more effective task-specific models.

4. Scope

Implement and compare different knowledge distillation methods (e.g., logit distillation, attention distillation) for generative language models.

Evaluate the distilled models on one or more NLP tasks (e.g., text classification, named entity recognition, question answering).

Analyse the trade-offs between model performance, computational efficiency, and data privacy when using distilled knowledge.

5. Potential Outcomes

Identify effective knowledge distillation techniques for generative language models.

Quantify the performance and efficiency gains of distilled models compared to the original large models.

Provide insights into the potential of knowledge distillation for deploying generative AI models in resource-constrained environments.

6. Primary Research Plan

An experimental strategy that applies, assesses, and analyses knowledge distillation methods.

Tasks and Timeline (13 weeks):

Week 1-2: Literature Review and Dataset Preparation

- Carry out an extensive review of the literature on methods for knowledge distillation for generative language models.
- Locate and prepare pertinent datasets for natural language processing (NLP) tasks (e.g., question answering, named entity recognition, and text categorization).

Week 3-5: Implementation of Knowledge Distillation Techniques

- Use various techniques for knowledge distillation (such as logit and attention distillation) for generative language models such as GPT-3 and BERT.
- Provide a framework for reducing large-scale models' knowledge to smaller, task-specific models.

Week 6-8: Evaluation and Benchmarking

- Use suitable assessment measures to assess the distilled models' performance on certain NLP tasks (e.g., accuracy, F1-score, BLEU score).
- Examine how well the distilled models perform in comparison to the huge original models and other baseline models.
- Compare the distillation models' computational efficiency (e.g., inference time, memory consumption).

Week 9-10: Analysis and Interpretation

- Using distilled knowledge, evaluate the trade-offs between computing efficiency, data privacy, and model performance.
- Examine how various distillation methods and hyperparameters affect the effectiveness and performance of the model.
- Analyse the findings and determine which knowledge distillation methods work best for generative language models.

Week 11-12: Documentation and Reporting

- Keep records of the experimental design, findings, and research technique.
- Write a thorough report that includes all the study's conclusions, analysis, and new information.
- List any possible drawbacks and suggest areas for further study.

Week 13: Finalization and Presentation

- Complete the report and make a presentation that includes a summary of the research findings.
- Inform colleagues and teachers about the research findings and insights.

Assumptions and Considerations:

The ability to train and assess big language models using computer resources (such as GPUs).

Large language models that have already been trained (like GPT-3 and BERT) and the matching distillation methods that go along with them.

Considering the task's complexity and probable difficulties, reasonable estimates about the quantity of work that can be completed in the allotted time.

Freedom to change the plan and parameters as needed in response to unforeseen events or surprising results discovered during the study process.

7. Ethical Considerations

Several significant ethical considerations for the chosen topic "Knowledge Distillation from Generative Language Models" should be kept in mind:

Data Privacy and Security:

GPT-3 and BERT, two large language models, are trained using enormous volumes of internet scraped data, some of which may contain private or protected content. Extracting information from these models may unintentionally encourage IP infringements or privacy violations. Anonymization, appropriate data filtering, and adherence to data protection laws like GDPR are essential.

Bias and Fairness:

Biases in society that are present in training data can be inherited and amplified by language models. Particularly in delicate fields like healthcare or finance, knowledge distillation may transfer these biases to the smaller distilled models, producing unfair or biassed results. During distillation, rigorous bias evaluation and mitigation strategies should be used.

Transparency and Accountability:

Large language models frequently have opaque black boxes inside of them, making it difficult to comprehend and explain their behaviour. Distilling knowledge might make decision-making even more difficult, which would raise questions about accountability and transparency, especially in high-stakes applications.

Misuse and Malicious Applications:

Although increasing efficiency and accessibility are the main objectives, it is possible that distilled language models may be abused to produce hate speech, false information, or other damaging content. To reduce these threats, strong security measures, content filtering, and conscientious deployment procedures are required.

Environmental Impact:

The computational demands of training big language models result in high energy consumption and carbon emissions. Although the goal of distillation is to produce more efficient models, the process's overall environmental impact should be considered and reduced whenever feasible.

Legal and Licensing Considerations:

Certain extensive language models, such as GPT-3, are licenced under specific conditions and are proprietary. Understanding these models and drawing conclusions from them may give rise

to legal questions around intellectual property rights and licence compliance. Adherence to legal frameworks and proper due diligence are crucial.

Throughout the study process, it is essential to take a responsible and proactive stance to address these ethical problems. This could entail working with ethics boards, carrying out thorough testing and audits, putting strong safety measures in place, and abiding by set ethical standards and norms for the creation and application of AI.

8. Literature Review

In recent years, there has been a substantial increase in interest in the concept of knowledge distillation, especially from big generative language models like BERT and GPT. To facilitate broader deployment and accessibility, this technique attempts to transfer knowledge from these computationally costly and resource-intensive models to smaller, more efficient task-specific models.

Various methods of knowledge distillation for generative language models have been investigated in several studies. MiniLLM, a technique that reduces huge language models into smaller models by using reverse Kullback-Leibler divergence as the goal, was proposed by Gu et al. (2024). Tang et al. (2019/2019) presented a method that effectively extracts task-specific knowledge from BERT and incorporates it into basic neural networks for a range of NLP applications.

Even though this research has contributed significantly, there are still some points of contention and possible gaps. For example, the efficacy of various distillation methods may differ for various tasks, domains, and model designs. To achieve sequence-to-sequence automatic speech recognition (ASR), Futami et al. (2020) concentrated on extracting knowledge from BERT, emphasising the necessity for task-specific distillation techniques.

Prominent research in this field now includes the works of Jiao et al. (n.d.) and Sanh et al. (2020), who proposed TinyBERT and DistilBERT, respectively. These studies indicate the trade-offs between model performance and efficiency, highlighting the possibility of extracting knowledge from big pre-trained models such as BERT for tasks related to natural language processing.

Still, there are some research gaps that might serve as launching pads for additional study. An area of potential interest is the assessment of distillation methods in a wider variety of natural language processing applications, like dialogue systems, machine translation, and summarization. Additionally, as evidenced by the ethical issues raised in a few of the publications (e.g., Beyer et al., 2022; Gou et al., 2021), the effect of distillation on model fairness and bias mitigation continues to be an understudied subject.

Investigating multi-stage or hierarchical distillation techniques, in which knowledge is distilled through several intermediate models, may be another fruitful avenue. These techniques may enable more effective and focused knowledge transfer. Moreover, as proposed by Mukherjee and Awadallah (n.d.) and Sun et al. (n.d.), the combination of distillation approaches with other model compression or acceleration methods, including quantization or pruning, may result in more comprehensive solutions for effective model deployment.

Regarding effective methodologies, the research conducted by Sanh et al. (2020) and Jiao et al. (n.d.) on DistilBERT and TinyBERT, respectively, offers significant perspectives on extracting knowledge from extensive pre-trained models such as BERT for tasks related to natural language comprehension. Their methods and assessment frameworks could be used as a basis for creating new distillation approaches or modifying current ones to fit other requirements or model designs. Furthermore, Gou et al.'s survey study from 2021 offers a thorough summary of knowledge distillation methods, making it a valuable tool for comprehending the state-of-the-art now and determining future research avenues.

Overall, even though knowledge distillation for generative language models has advanced significantly, there is still need for more study and creativity. Examining the literature from a critical and evaluative standpoint emphasises the necessity for multi-stage or integrated distillation procedures, thorough evaluation across a range of tasks and domains, and task-specific distillation methodologies. As these condensed models are used in practical applications, it should also be a top focus to address ethical issues like fairness and bias reduction.

References

1. Gu, Y., 1, Dong, L., Wei, F., Huang, M., 1, The CoAI Group, Tsinghua University, & Microsoft Research. (2024). MiniLLM: Knowledge Distillation of Large Language Models. arXiv Preprint. https://arxiv.org/pdf/2306.08543.pdf (Gu et al., 2024)
2. Futami, H., 1, Inaguma, H., 1, Ueno, S., 1, Mimura, M., 1, Sakai, S., 1, Kawahara, T., 1, & Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan. (2020). Distilling the knowledge of BERT for Sequence-to-Sequence ASR. In arXiv [Journal-article]. https://arxiv.org/pdf/2008.03822.pdf (Futami et al., 2020)
3. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Kolesnikov, R. A. A., & Google Research, Brain Team. (2022). Knowledge distillation: A good teacher is patient and consistent. In arXiv:2106.05237v2 [cs.CV] 21 Jun 2022. https://arxiv.org/pdf/2106.05237.pdf (Beyer et al., 2022)
4. Suˇc´Ik, S. (2020). Teacher-student knowledge distillation from BERT. In University of Edinburgh, Master of Informatics. https://project-archive.inf.ed.ac.uk/ug4/20201880/ug4_proj.pdf (Suˇc´Ik, 2020)
5. Mukherjee, S., & Awadallah, A. (n.d.). DISTILLING BERT INTO SIMPLE NEURAL NETWORKS WITH UNLABELED TRANSFER DATA *. https://arxiv.org/pdf/1910.01769.pdf (Mukherjee & Awadallah, n.d.)
6. Sun, S., Cheng, Y., Gan, Z., Liu, J., & Microsoft Dynamics 365 AI Research. (n.d.). Patient knowledge distillation for BERT model compression. Microsoft Dynamics 365 AI Research. https://arxiv.org/pdf/1908.09355.pdf (Sun et al., n.d.)
7. Sanh, V., Debut, L., Chaumond, J., Wolf, T., & Hugging Face. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Journal-article]. Hugging Face. https://arxiv.org/pdf/1910.01108.pdf (Sanh et al., 2020)
8. Jiao, X., 1, Yin, Y., 2, Shang, L., 2, Jiang, X., Chen, X., Li, L., Wang, F., 1, Liu, Q., Key Laboratory of Information Storage System, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Huawei Noah's Ark Lab, & Huawei Technologies Co., Ltd. (n.d.). TinyBERT: Distilling BERT for Natural Language Understanding. arXiv Preprint. https://arxiv.org/pdf/1909.10351.pdf (Jiao et al., n.d.)
9. Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., Lin, J., & University of Waterloo. (2019). Distilling Task-Specific Knowledge from BERT into Simple Neural Networks

[Journal-article]. University of Waterloo. https://arxiv.org/pdf/1903.12136.pdf (Original work published 2019) (Tang et al., 2019/2019)

10. Gou, J., Yu, B., Maybank, S. J., Tao, D., UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia, & Department of Computer Science and Information Systems, Birkbeck College, University of London, UK. (2021). Knowledge distillation: a survey. https://www.dcs.bbk.ac.uk/~sjmaybank/KD_Survey-arxiv.pdf (Gou et al., 2021)