Project Proposal: Ensemble Domain-Specific Knowledge Distillation

Camden Shultz, Kevin Kim, Sara Ren

Motivation:

Even with recent advances in natural language processing and machine learning, large language models (LLMs) tend to suffer from two big problems: they're too big and they hallucinate. Especially as LLMs become more and more widespread in the general population, there will be a greater need for more memory efficient and accurate models, especially those that are capable of answering more specific questions in many different topics. We aim to mitigate the issues of overly large size and hallucinations with one solution: domain-specific knowledge distillation. Knowledge distillation is the process of training a smaller model (the "student") on the input-output pairs of the larger (the "teacher") model. Domain specific distillation requires reducing both the parameter size and vocabulary size while fine tuning to restricted domains of knowledge, such as law, medicine, math, or computer science, which allows models to perform more accurately in domain-specific answering than domain-agnostic distillation. Using knowledge distillation, we hope to construct an ensemble of smaller and more efficient domain-specific and domain-agnostic models that performs similarly to a large model on domain-agnostic tasks, while also reducing hallucinations compared to the larger models for specific tasks.

Relevant existing literature:

Adapt-and-Distill: Developing Small, Fast and Effective Pre-Trained Language Models for Domains[1]

This article describes the results of four methods of domain specific knowledge distillation. Of the four methods, the "adapt-and-distill" method achieves state-of-the-art results. This method essentially fine-tunes both the student and teacher models to a specific domain before distilling from the larger model to the smaller model.

Mixture-of-Experts with Expert Choice Routing[2]

The paper proposes a routing mechanism for sparsely-activated mixture-of-experts models that allows each expert to select the top-k tokens instead of tokens selecting experts. It was noted that this method achieves better load balancing across experts and better training efficiency compared to previous methods. Inspired by this, we plan to develop an attention mechanism across the ensemble of student models, attending to each model's strengths ad hoc.

Domain-specific knowledge distillation yields smaller and better models for conversational commerce[3]

It was shown that distilled domain-specific models on average performed 2.3% better than the generic distilled model. It was also shown that while most adaptations of BERT fine-tune encoder weightings, improvements from distillations on target data mostly remain even when encoder weights are frozen during student model training. Encoders that were fine-tuned still performed better than domain-specific distillation, however.

Knowledge Distillation Transfer Sets and their Impact on Downstream NLU Tasks[4]

The article explores the impact of using domain-specific versus generic data sets in the knowledge distillation process. The study shows that models distilled using only domain-specific data perform better on their target tasks than those distilled on a mix of generic and domain-specific data. The paper also investigates the impact of distillation after the teacher model has been pretrained (on generic data) and before fine tuning. It was found that first fine tuning and then distilling produced the best results.

---

[1] Yao et al.
[2] Zhou et al.
[3] Howell et al.
[4] Peris et al.

Project Plan

Hypothesis & expected outcome:

We assume that a lot of large language models are currently filled with redundant parameters. By using a combination of domain-specific and domain-agnostic distillation to train smaller and more parameter-efficient language models, we can mitigate the issues associated with large language models, such as their size and tendency to hallucinate, while maintaining performance on domain-agnostic tasks. Recent demonstrations, such as GPT-4 and large language models have shown success in leveraging model capacity to place among higher percentiles in standardized tests. We believe that an ensemble of domain-specific knowledge distilled models can achieve comparable, if not higher, performance on domain-specific questions while maintaining comparable results on domain-agnostic questions and will have an overall lower memory footprint.

Experiments:

1. **Models:** We will use GPT as our teacher model, and if time permits, OPT.
    a. Inspired by our reading, we intend to use an attention-layer over the ensemble for gating queries. Other controllers, including simple classifiers, may also be explored.
2. **Benchmark:** Evaluation of large teacher models on standardized benchmarks and domain-specific knowledge/Q&A.
3. **Distillation Baseline:** Train a single student model instead of an ensemble of student models, evaluating the effects of model compression and distillation.
4. **Expert Student Ensemble:** Evaluation of ensemble of domain-specific student learners on standardized benchmarks and domain-specific Q&A.
5. **Ablation Studies:**
    a. **Global Vocabulary:** Train student models using input-output pairs from the dataset, without restricting vocabulary to a specific domain. This would evaluate the impact of vocabulary trimming on the student model.
    b. **Non-Ensemble:** Train a single student model instead of an ensemble of student models, evaluating the effects of model compression and distillation.

Success Metrics: Evaluate and compare to accuracy of LLMs on standardized tests meant for human test-takers (Olympiad, Bar exam), compare size of ensemble model to LLM.[5]

Datasets:

We think that StackExchange will be an extremely valuable dataset,[6] since it is a huge repository (24 million questions and 35 million answers) of labeled, carefully moderated question/answer pairs that are separated by topic. For domain specific knowledge, we can additionally refer to textbook/wikipedia materials (or other publicly available encyclopedic data) that are available online. The goal is to fit strongly to a specialized corpus (*forum*) for each distilled model, so any dataset with a large amount of truth will work. That is, our transfer set will be task-specific, rather than task-agnostic.

Halfway Milestone:

In our halfway milestone, we hope to have a pipeline for a single downstream task complete that allows us to have a specialized distilled model on Math Olympiad questions. By having this pipeline complete, we can then generalize to other tasks with relative ease (where the only limiting factor will be training time) and focus on evaluation as well as ablation.

---

[5] OpenAI.
[6] Howell et al.

References

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-Distill: Developing Small, Fast and Effective Pretrained Language Models for Domains. *arXiv preprint arXiv:2106.13474.*

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022. Mixture-of-Experts with Expert Choice Routing. *arXiv preprint arXiv:2202.09368.*

Kristen Howell, Jian Wang, Akshay Hazare, Joseph Bradley, Chris Brew, Xi Chen, Matthew Dunn, Beth Ann Hockey, Andrew Maurer, and Dominic Widdows. 2022. Domain-specific knowledge distillation yields smaller and better models for conversational commerce. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5),* pages 151-160, Online. Association for Computational Linguistics.

Charith Peris, Lizhen Tan, Thomas Gueudre, Turan Gojayev, Pan Wei, and Gokmen Oz. Knowledge distillation transfer sets and their impact on downstream NLU tasks. 2022. In *EMNLP 2022*.

OpenAI. GPT-4 Technical Report. 2023. *arXiv preprint arXiv:2303.08774*