Role of Small Models in the LLM Era

Collaboration

LLMs Enhance SMs

- **Knowledge Distillation**
  - Black-Box Knowledge Distillation: involves generating a distillation dataset through the teacher LLM, which is then used for fine-tuning the student model.
  - White-Box Knowledge Distillation: using internal states of the teacher model in the training process of the student model.
- **Data Synhesis**
  - Generating a dataset from scratch using LLMs, in an unsupervised manner, followed by training a task-specific SM on the synthesized dataset.
  - Leveraging LLMs solely to generate labels rather than the entire training dataset.
  - Using LLMs to modify existing data points, thereby increasing data diversity, which can then be directly used to train smaller models, Eg: paraphrase or rewrite texts.
- **Data Curation**
  - Small model can be trained specifically to evaluate text quality, enabling the selection of high-quality subsets.
  - Perplexity scores can be calculated by a SM to select data that is more likely to be of high quality.
  - SMs can be used as classifiers to evaluate instruction data based on quality, coverage, and necessity.

SMs Enhance LLMs

- **Weak-to-Strong Paradigm**
  - LLMs can be fine-tuned on labels generated by a diverse set of specialized SMs, enabling the strong models to generalize beyond the limitations of their weaker supervisors.
  - SMs can also collaborate with LLMs during the inference phase to further enhance alignment.
- **Efficient Inference**
  - Model Cascading
  - Model Routing
  - Speculative Decoding: The auxiliary SM can quickly generates multiple token candidates in parallel, which are then validated or refined by the LLM.
- **Evaluating LLMs**
  - Model-based evaluation approaches can use smaller models to assess performance. Eg: BERT Score.
- **Domain Adaptation**
  - Black-Box Adaptation involves using a domain-specific SM to guide LLMs toward a target domain by providing textual relevant knowledge.
  - White-Box Adaptation typically involves fine- tuning a SM to adjust the token distributions of frozen LLMs for a specific target domain.
- **Retrieval Augmented Generation**
  - Retrievers based on SMs can be used for enhancing generations, Eg ColBERT
- **Prompt-based Learning**
  - SMs can be employed to enhance prompts, thereby improving the performance of larger models.
  - SMs can be used to verify or rewrite the outputs of LLMs, thereby achieving performance gains without the need for fine-tuning.
- **Deficiency Repair**
  - SMs can leverage contrastive decoding to reduce repetition, hallucinations in LLMs.
  - Specialized fine-tuned SM can be used to address some of the shortcomings of the larger model.

Competition

- **Computation-constrained Environment**
  - Small models are increasingly valuable in scenarios where computational resources are limited.
- **Task-specific Environment**
  - Small tree-based models can achieve competitive performance compared to large deep learning models for tabular data.
  - Fine-tuning SMs on domain-specific datasets can outperform general LLMs.
  - SMs can be particularly effective for tasks such as text classification, phrase representation, and entity retrieval.
- **Interpretability-required Environment**
  - Generally, smaller and simpler models offer better interpretability compared to larger, more complex models.