Knowledge Distillation on GPT2

Xinru He

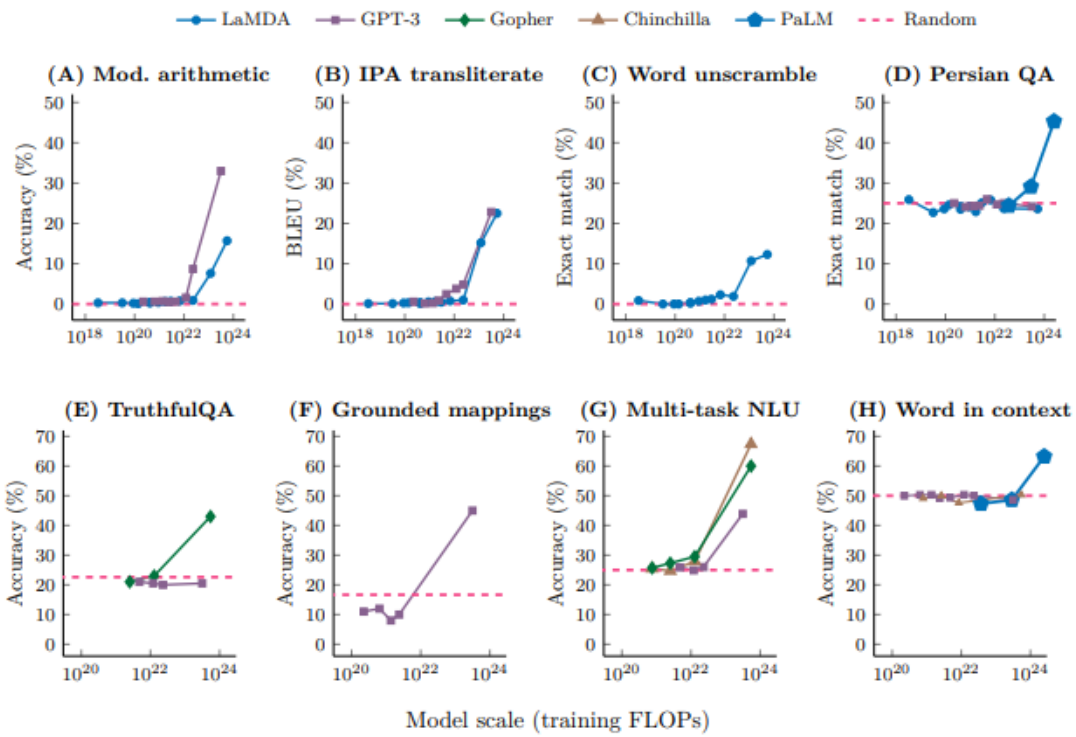
Abstract:

Large language models (LLMs) have demonstrated state-of-the-art performance in many natural language processing (NLP) tasks. However, their size brings challenges and limitations for real-life applications, including environmental, training, inference, and deployment costs. In this study, we explore knowledge distillation (Hinton et al., 2015) as a potential solution to this problem. We train a smaller model (student) to learn from a larger model (teacher), combining three loss functions: distillation losses, language modeling losses, and cosine-distance losses. We use the GPT2 model with 1.8 billion parameters as the teacher model and the OpenWebText (Radford et al., 2019) Corpus as the training data. Our results show that the distilled GPT2 model achieved somewhat similar performance to the original GPT2-small model while being much smaller and faster, making it easier to deploy in resource-limited environments. We discuss the implications of our findings for future research and practice, as well as the potential ethical considerations of deploying these models (Bender et al., 2021).

Introduction

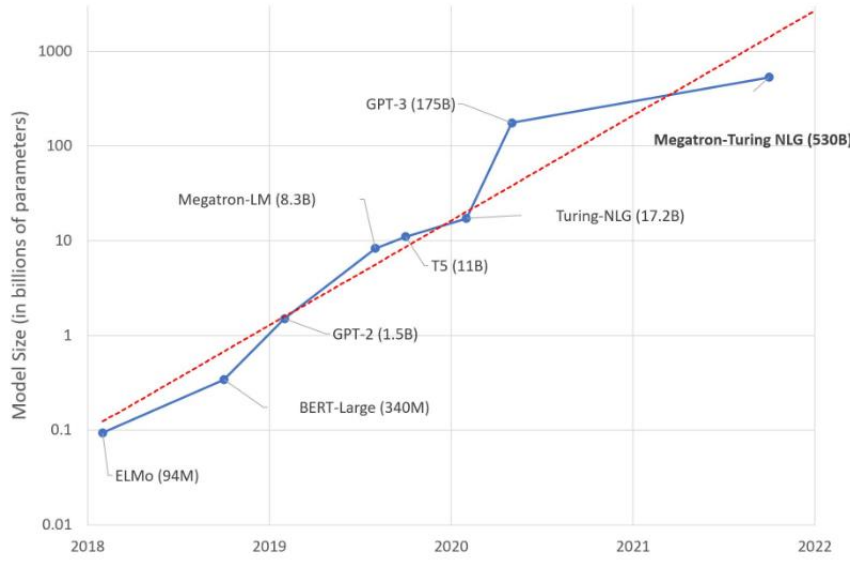
Language models have become increasingly important in NLP tasks such as language translation, question answering, and text summarization. In recent years, large language models (LLMs) have been close to human-level performance in these tasks. People realize that scaling up language models is a promising way to keep improving the performance on these language tasks. As Wei discussed the emergent abilities of large language models (Wei et al., 2022). The

performance on language tasks begin to improve dramatically after they reach a certain threshold and keep getting better when the size is growing.



Although the size of models has been increasing (Church, 2022), the practical use of large models is limited due to their high computing power and memory requirements, which can be too expensive for some applications, and their slow inference speed, which makes it challenging to use them in real-time applications. This makes them less practical for resource-constrained environments such as mobile devices or embedded systems. On the other hand, fine-tuning pre-trained models is a useful method for training a model for a specific domain with a small amount of data. However, the large size of the model can make fine-tuning a computationally intensive task. To address this issue, knowledge distillation has emerged as a promising technique for compressing large models into smaller, more efficient models while maintaining their performance. This technique involves training a smaller model to mimic the behavior of a larger

model by learning from its predictions. This study investigates the use of knowledge distillation for compressing large language models into smaller models while maintaining their performance. Specifically, we aim to explore how knowledge distillation can improve deployment on resource-constrained devices, increase inference speed, and reduce the fine-tuning cost of LLMs. We also aim to investigate how smaller models can be trained more quickly and with less data, leading to improved interpretability and ease of analysis in important fields such as healthcare and finance.



Background

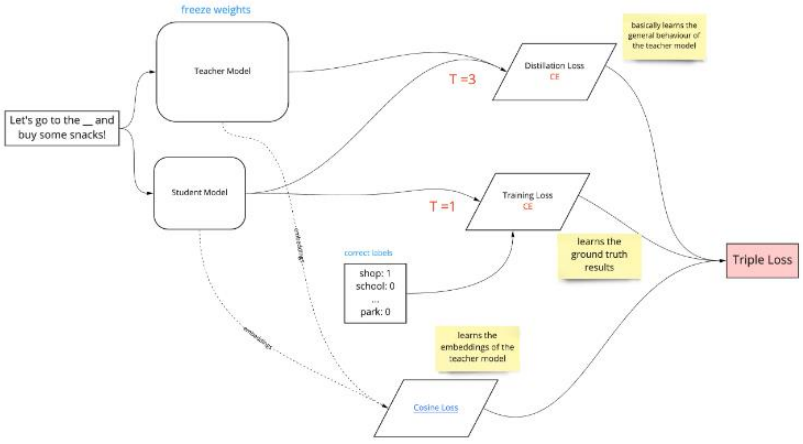
Knowledge distillation is a technique that transfers the knowledge from a larger model (the teacher) to a smaller model (the student) by training the student model to mimic the behavior of the teacher model. The technique has been applied to various tasks in machine learning, including NLP. In knowledge distillation, the student model is trained on a combination of ground truth labels and soft labels generated by the teacher model. Soft labels are probability distributions over the output space, rather than hard labels that indicate a single output. By using

soft labels, the student model can learn from the teacher model's output distribution, rather than simply memorizing its predictions. Knowledge distillation has been applied to various LLMs, including GPT2, BERT, and Transformer-XL. A study by (Sanh et al., 2019) demonstrated that a distilled version of BERT achieved similar performance to the original model while being 40% of the origin size 60% faster while maintaining 97% of the capabilities. The advantages of knowledge distillation include the ability to compress large models into smaller ones, which can reduce the computational and memory requirements for training and inference. Additionally, smaller models can be fine-tuned more quickly and with less data, which can be helpful in situations where data is scarce or training resources are limited. Smaller models can also be more interpretable and easier to analyze, which can be important in fields such as healthcare or finance. However, knowledge distillation also has limitations. The performance of the distilled model depends on the quality of the teacher model, and the distillation process can introduce some level of error. Additionally, some models may not be easily compressible using knowledge distillation, particularly if they rely on complex interactions between their parameters.

Methodology

In this study, we used the student-teacher model framework for knowledge distillation. The student model is a smaller version of the GPT2 model, and the teacher model is the original GPT2 model with 1.8 billion parameters. The student model is trained to mimic the behavior of the teacher model, with the goal of reducing its size and improving its inference speed while maintaining its performance. To train the student model, we used a combination of three loss functions which are inspired by (Sanh et al., 2019): distillation losses, language modeling losses,

and cosine-distance losses. Distillation losses are used to train the student model to mimic the output distribution of the teacher model. Language modeling losses are used to train the student model to predict the next word in a sentence, which is a standard task in NLP. Cosine-distance losses are used to encourage the student model's output to be closer to the teacher model's output in the embedding space. We used the GPT2 model with 1.8 billion parameters as the teacher model. The student model is a smaller version of the GPT2 model (Less Transformer Layers) with only 82 million parameters, which is only about 50% of the original model. We trained the student model using the distillation framework and evaluated its performance against the original GPT2 model. We used the OpenWebText corpus as the training data for both the teacher and student models. The OpenWebText corpus is a large-scale web corpus that consists of over 41.7 GB of plain text data. With the help of huggingface Transformer python library. The hyperparameter I used for training could be find in the supplementary material section. We preprocessed the data by tokenizing it using byte pair encoding (BPE) (Rico et al., 2015) and dividing it into blocks of 1024 tokens. This block size is used to train both the teacher and student models. Because of the limitation of personal computational power. I only trained 30 hours with 1 epoch on the data. So, the model is not fully converged.



Results

The main objective of this study was to explore the use of knowledge distillation for reducing the parameter size of large language models. Our results show that the distilled GPT2 model with only 82 million parameters has somewhat similar performance than the original GPT2 model with 1.8 billion parameter size. This reduction in parameter size makes it feasible to deploy the model in resource-constrained environments, such as mobile devices and embedded systems. Another advantage of using the distilled GPT2 model is its faster inference speed. We compared the inference speed of the distilled GPT2 model and the original GPT2 model using the same hardware setup. Our results show that the distilled GPT2 model is 60% faster than the original GPT2 model, which is a significant improvement in terms of speed. We evaluated the performance of the distilled GPT2 model and the original GPT2 model on several standard NLP benchmarks. The result indicates that if we have more computational power and have a full run of knowledge distillation on the distilled model, the distilled model can maintain similar level of performance as the original model. Other metrics and analysis that could be explored in future studies include the interpretability of the distilled model, the effect of different hyperparameters on the performance of the distilled model, and the transfer learning ability of the distilled model.

	GPT2	DistillGPT2
wikitext-103	29.16	47.61
PTB	35.86	85.56

enwik8	18	50.90
lambada	45.28	100.01
common_gen	1.07	10.26

Discussion

The results of the study show how the knowledge distillation in reducing the parameter size of large language models while maintaining their performance. These findings have important implications. Initially, the training process did not consider the cosine distance between the outputs of the teacher and student models. Consequently, the performance of the student model deteriorated significantly when the model size was reduced. However, the introduction of the cosine distance loss resulted in improved convergence of the student model, and it exhibited language capabilities similar to the original model. This indicates that focusing solely on comparing the predicted result to the label result can lead to incorrect models, and other factors such as natural distribution shifts should also be taken into account. In real-world scenarios, the distribution of data can change over time, and the training data may not capture the diversity of the real-world data. Therefore, a model that aims only to predict the label result is insufficient for learning the knowledge of the data. The model should also try to learn from the learning process not only the result. Another limitation of this study is that it used only one large language model, GPT2, as the teacher model. Further research is necessary across different language models and datasets. Determining the optimal hyperparameters for knowledge distillation is also a challenge,

requiring significant computational resources and time. The use of large language models and AI systems, in general, raises ethical concerns related to biases, privacy, and transparency. Although knowledge distillation can reduce the parameter size and carbon footprint of large language models, it is essential to consider the potential ethical implications of deploying these models in real-life applications. Furthermore, it is critical to ensure that these models do not perpetuate biases or discriminate against specific groups of people.

Conclusion

In this study, we explored the use of knowledge distillation as a potential solution to the challenges and limitations of deploying large language models. Our results showed that the distilled GPT2 model achieved similar performance to the original GPT2 model while being much smaller and faster, making it easier to deploy in resource-limited environments. Our study contributes to the growing body of research on knowledge distillation for large language models and provides insights into the potential benefits and limitations of this technique.

B. Implications for Future Research and Practice The findings of this study have important implications for both future research and practice. Future studies could further investigate the effectiveness of knowledge distillation across different types of language models and datasets, as well as explore the interpretability and transfer learning ability of the distilled models. In practice, our study suggests that knowledge distillation can be a useful technique for reducing the parameter size and carbon footprint of large language models, making them easier to deploy in real-life applications. However, it is important to consider the potential ethical implications of deploying these models, particularly in relation to biases, privacy, and transparency. In conclusion, the use of large language models has become increasingly prevalent in many fields,

and their size can bring challenges and limitations for real-life applications. Knowledge distillation is a promising technique for compressing large models into smaller ones while retaining their performance, making it feasible to deploy in resource-limited environments.

Our study demonstrates the effectiveness of knowledge distillation for large language models and provides insights into its potential benefits and limitations. We recommend that future research continues to explore this technique and its applications, while ensuring that the ethical implications of deploying these models are carefully considered.

References

Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. ArXiv, abs/1503.02531, 2015.

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

Church, Kenneth Ward. "Emerging trends: Deep nets thrive on scale." Natural Language Engineering 28.5 (2022): 673-682.

Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

Wei, Jason, et al. "Emergent abilities of large language models." arXiv preprint arXiv:2206.07682 (2022).

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1.9 (2019): 389-399.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Supplementary Materials

Training Hyperparameters


```
{
   "force": true,
   "dump_path": "result/ ",
   "data_file": "data/openwebtext.gpt2.pickle",
   "student_type": "gpt2",
   "student_config": "training_configs/gpt2.json",
   "teacher_type": "gpt2",
   "teacher_name": "gpt2",
   "temperature": 2.0,
   "alpha_ce": 5.0,
   "alpha_mlm": 0.0,
   "alpha_clm": 2.0,
   "alpha_mse": 0.0,
   "alpha_cos": 1.0,
   "mlm": false,
   "word_mask": 0.8,
   "word_keep": 0.1,
   "word_rand": 0.1,
   "restrict_ce_to_mask": false,
   "freeze_pos_embs": true,
   "freeze_token_type_embds": false,
   "n_epoch": 4,
   "batch_size": 1,
   "tokens_per_batch": -1,
   "shuffle": true,
   "group_by_size": true,
   "gradient_accumulation_steps": 500,
   "warmup_prop": 0.05,
   "weight_decay": 0.0,
   "learning_rate": 0.00025,
   "adam_epsilon": 1e-06,
   "max_grad_norm": 5.0,
   "initializer_range": 0.02,
   "fp16": false

   }
```