Name: Abdullah

Matriculation Number: 7075730

Title: Overtrained Language Models Are Harder to Fine-Tune

Authors: Jacob Mitchell, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, Aditi Raghunathan

Summary

This paper challenges a widely held belief in AI: that training language models on more and more data always makes them better, often summarized by the idea that scaling laws alone are sufficient for improved model performance. While longer pre-training does improve a model's raw capabilities, the authors show that it can actually hurt performance after fine-tuning, which is the step where models are adapted to specific tasks like instruction following or multimodal reasoning. In multiple real-world experiments, models that were trained on more data performed worse after fine-tuning than models trained on less data. For example, an OLMo-1B model trained on 3 trillion tokens performed 2–3% worse on standard benchmarks after instruction tuning than the same model trained on 2.3 trillion tokens. Crucially, the base models (before fine-tuning) did improve with more pre-training; the performance drop only appeared after fine-tuning.

The research identifies a key mechanism behind this effect, called "progressive sensitivity". As a model is pre-trained longer, It learns more features and becomes more specialized, but it also becomes more sensitive to parameter changes. Fine-tuning, by design, modifies the model's parameters. The results show that the same size change (for example, the same learning rate or amount of noise) causes more damage to a heavily pre-trained model than to a less-trained one. That's when "catastrophic overtraining" occurs. The degradation (or "forgetting") caused by this progressive sensitivity eventually grows faster than the base model's intrinsic improvement from extended pre-training, even with careful hyperparameter tuning. In essence, a highly overtrained model becomes brittle: it performs better on its general pre-training task, but any attempt to specialize it through fine-tuning causes a larger and more damaging distortion of its general knowledge.

The authors emphasize that a singular focus on maximizing pre-training performance is therefore counterproductive. Researchers face an inherent trade-off: while adjusting the fine-tuning process (for example, by using smaller learning rates) can delay the onset of catastrophic overtraining, it can also limit the model's capacity to learn the new, desired task. The key takeaway is that future model design must rethink pre-training token budgets and optimization strategies to balance raw capability with downstream adaptability.