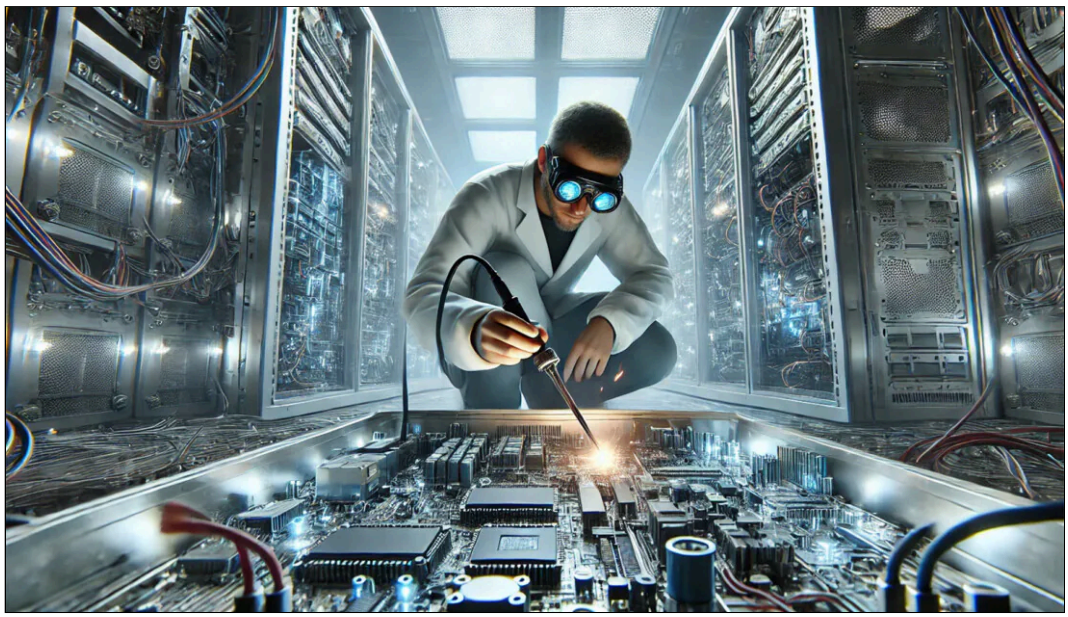
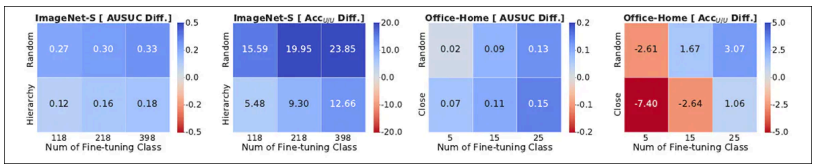
The Damage From Fine-Tuning an AI Model Can Easily Be Recovered, Research Finds

Originally published at https://www.unite.ai/the-damage-from-fine-tuning-an-ai-model-can-easily-be-recovered-research-finds/



New research from the US indicates that fine-tuning an AI foundation model on your own data does not need to reduce or impair the functionality of the original model – and that a relatively simple fix can not only restore the capabilities of the original model, but actually *improve* the quality of the output that you're trying to get the (already trained) model to produce.



Performance gains on diverse models with the authors' new post-training calibration. Further details later in the article. Source: http://export.arxiv.org/pdf/2409.162

The implications for this are significant, not only for the tech giants whose attentions are converging on the financial rewards of renting out generative systems 'as-a-service', but also the growing number of 'cord-cutter' hobbyists who download and customize open source models, so that they can access personalized AI writing and image/video generation systems more cheaply – and with fewer restrictions.

The authors of the paper are not afraid to show their enthusiasm for the potential of their method, which makes apparently significant advances on the 2023 submission *Holistic Transfer: Towards Non-Disruptive Fine-Tuning with Partial Target Data* (co-authored with many of the contributors to the new paper).

They state:

> *'The [findings] are encouraging and have profound implications! They imply that a simple post-processing calibration can potentially address the fine-tuned model's inferior accuracy on the absent classes, bringing back the pre-trained model's capability while unveiling the improved feature quality over all classes.'*

We'll take a look at the new work shortly. First, let's see what problem it is aiming to solve.

Why It Matters

The first wave of widespread fine-tuning occurred in the wake of the release of Stability.ai's Stable Diffusion text-to-image model in August 2002. The early models, trained on a subset of the hyperscale LAION dataset, were made available for anyone to download.

However, users who wanted to insert *specific* content (such as their own identities, art styles, or the representation of celebrities) into the extraordinary generative qualities of Stable Diffusion were required to turn to techniques such as DreamBooth – an extrapolation of a Google Research customization method, which allowed the user to train new data into the freely-available model, via fine-tuning.

Examples of the user process for Google's official DreamBooth implementation from 2022. The user curates a small selection of images and chooses a unique name (one that Stable Diffusion does not have in its training data) in text-prompts from the fine-tuned model. Source: https://dreambooth.github.io/

In this way, it was possible to get a copy of the model that was very good at creating a particular person, or a custom art style, but which was *now 'compromised' for more general usage*.

This meant that if you wanted to fine-tune Stable Diffusion so that it could accurately depict three different people, you inevitably had to create *three different models*, each around 2-4GB, or more.

Any attempt to fine-tune these models *a second* time would not only degrade general performance of the model even further, but would adversely affect output from the previous fine-tuning session.

In any case, celebrity DreamBooth models would soon proliferate on the internet, convening primarily at the civit.ai domain. Eventually, less onerous methods such as Low-Rank Adaptation (LoRA) overtook fine-tuning in popularity (though whether LoRA output is as effective as a full fine-tune remains contentious, and NVIDIA NVDA 8.74% has since open-sourced an apparently more effective approach called DoRA).

A LoRA falls under the category of Parameter-Efficient Fine-Tuning (PEFT), which only influences a subset of the model's trained parameters.

Some users wanted to change the fundamental nature of the open sourced Stable Diffusion checkpoints, by fine-tuning them on many thousands of images.

This, effectively, produced an alternate foundation model, dedicated to whatever domain the user was trying to train (such as a particular art style). For this purpose, 'lightweight' methods such as LoRA were likely to be less effective, since the weights of the model needed a *severe* bias towards the new training data.

Local Chat

With the recent upsurge of interest in Large Language Models (LLMs), users wishing to avoid the growing outlets (and associated costs) of API-driven services such as ChatGPT, have increasingly started to download and fine-tune effective open source models like Llama 3, among many others.

Here too, LoRAs can be used instead of fine-tuning a full checkpoint. We have contended before that fine-tuning is a superior method for producing LLMs that are adapted to the specific user's needs. Though fine-tuning can have greater hardware requirements and may take longer, it offers a deeper generalization of the novel data that the user wants the model to assimilate.

The trouble with fine-tuning is that it's a destructive process that can't be incrementally trained on additional data later, as we noted above.

The features and biases being injected into the model apparently upset the original balance of weights in the dataset, meaning that the model is either excessively likely to reflect that user-contributed data, or will at least perform worse overall than the original foundation model (on tasks that are unrelated to the new data).

One can remedy this, to a certain extent, by freezing certain parts of the model during training; but this can lead to reduced general functionality, since the frozen part of the architecture may not generalize well to the newly fine-tuned data inside the model's latent space.

It would, therefore, be really great if there was some easier way to preserve the original capabilities of a fine-tuned model, while retaining the model's ability to produce output based on the fine-tuning data.

Such a development would be beneficial across the range of potential users, from hobbyists and early adopters using local LLMs and other types of generative model, up to FAANG-level (where a very expensive AI model could be improved iteratively and non-destructively, without the multi-million dollar expense of starting the training all over again with the additional data).

Post-Processing Calibration

This brings us back to the new paper, which is called *Fine-Tuning is Fine, if Calibrated*, and comes from 11 researchers across Ohio State University, the University of Wisconsin Madison, and the Rensselar Polytechnic Institute.

The researchers were attempting to find out exactly what gets damaged in a foundation model when it is fine-tuned. They have concluded that the only major difference between the 'before and after' model is that the logit scales across the fine-tuning classes and the original classes in the model exhibit a major discrepancy.

Logit links predict the probability of success in a logical regression process, converting the estimated values (which may be very precise) into a zero or a one.

The authors not only found that this deficit is almost casually reversible by a calibration technique, but that this *post facto* fix actually improves the quality of output for the fine-tuning data. Therefore, with this technique, you not only get the original capabilities of the foundation model, but you get a better integration of your own fine-tuned data.

(Though the paper does not examine the prospect, this technique implies that a model could be fine-tuned multiple times, and remain effective)

Discussing their findings in investigating model damage after fine-tuning, the authors state:

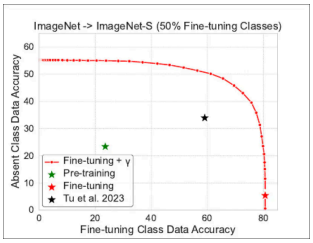
> *'To our surprise, we find that the fine-tuned model neither forgets the relationship among the other classes nor degrades the features to recognize these classes.*
>
> *'Instead, the fine-tuned model often produces more discriminative features for these other classes, even if they were missing during fine-tuning!*
>
> *'[What] really hurts the accuracy is the discrepant logit scales between the fine-tuning classes and the other [classes], implying that a simple post-processing calibration would bring back the pre-trained model's capability and at the same time unveil the feature improvement over all classes.'*

The authors have made the results of their tests for this theory reproducible in a GitHub repository.

They found that on investigation, the only part of the foundation model's architecture that is damaged in fine-tuning is the binary classifier, which misclassifies classes that are *absent* in the original model as fine-tuning classes.

The paper states*:

> *'[By] adding a calibration bias factor to all the absent classes' logits [4, 40], the fine-tuned model can successfully reclaim the absent class accuracy and obtain decent overall improvement in the downstream [domain].*
>
> *'The resulting performance even beats the strong baseline [Holistic Transfer – the paper on which this paper builds] in many of the benchmarks, including ImageNet and its variants [ImageNet, ImageNet-R(endition), ImageNet-S(ketch)], Office-Home, and VTAB,* without complicated training and hyperparameter setting*.'*



Results from the paper: a fine-tuned model that has had post processing calibration performed on it can, the authors state, outperform the state-of-the-art approach to the problem.

The authors classify the improved performance of a post-calibrated fine-tuned model as 'unexpected benign behaviors', and observe that when a basic Stochastic Gradient Descent (SGD) optimizer is used, a better result is obtained than with more popular current optimizers, such as Adam.

'Still,' they note *'with smaller enough learning rates and weight decay, the benign behaviors show up and hold.'*

Minor Repairs

To repair the logit discrepancies resultant from fine-tuning, the authors borrowed a technique from zero-shot learning, adding a constant factor to the logits of all the absent classes. This results in a new classification rule.

The authors note that this process 'promotes' the neglected absent classes to the same prediction quality of the fine-tuned classes, restoring original performance and improving the performance of the 'added' data at inference time.

	ImageNet-{R, S}			VTAB			Office-Home		
Metrics (%)	$\text{Acc}_{\mathcal{Y}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{S}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{U}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{Y}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{S}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{U}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{Y}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{S}/\mathcal{Y}}$	$\text{Acc}_{\mathcal{U}/\mathcal{Y}}$
Pre-trained	23.6	23.5	23.8	58.3	59.3	57.4	63.8	61.8	65.5
Fine-tuning	43.3	81.3	3.5	62.8	86.4	39.1	53.5	88.3	22.4
Tu *et al.* 48	47.5	62.1	23.1	63.8	83.9	43.5	72.0	78.2	66.6
Fine-tuning + γ_{ALG}	55.9	80.3	30.5	66.8	85.3	48.2	65.0	87.7	44.9
Fine-tuning + γ_{PCV}	57.1	60.1	54.0	57.4	47.1	67.8	72.2	82.3	63.1
Fine-tuning + $\gamma\star$	60.8	73.6	47.6	69.3	75.6	62.8	72.7	79.1	66.9
Oracle	71.1	72.4	69.8	80.6	79.8	81.3	82.1	81.2	82.9

In tests, the post-calibration technique restored performance to a diversity of fine-tuned models. The 'Oracle' indicated in the table refers to a fine-tuned classifier that also takes into consideration missing class data.

They observe further that post-processing calibration is 'potentially applicable to any model', and that methods that seek to maintain foundation model integrity via the freezing of layers (such as the classifier and the backbone) score poorly in comparison to their own proposed approach.

Conclusion

The findings from this collaboration appear significant. Training an AI model on a hyperscale dataset is an enormous commitment, analogous to the take-off of a passenger jet. Though training can be interrupted, and any damage mitigated by saving the current weights periodically (at considerable storage cost), to allow interruptions to training, there is relatively little one can do to alter the outcome after launch.

What's impressive about the work is that the researchers seem to have discovered a fundamental principle in general AI model training, and that their solution is surprisingly elegant.

The economic implications of being able to retain foundation model accuracy after fine-tuning are also significant. To date, the most common method of addressing the shortcomings of multi-million dollar models has been to filter output at inference time, or to control inference in order to avoid any Achilles heel evident in the model.

Additionally, such a technique could theoretically bring significant improvements to the capabilities of fine-tuned generative models at the consumer level, with the bonus of a boost in output quality.

** My conversion of the authors' inline citations to hyperlinks.*

First published Tuesday, October 1, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More