Andrew Taylor EN.705.743 11/11/25

Reading Assignment 11: LoRA

Q1.

LoRA improves over supervised fine-tuning in multi-domain learning by treating adaptation as a small, structured deviation instead of a full overwrite of the base model. In traditional SFT, each new task drags the model's parameters in a new direction, causing old skills to erode. LoRA sidesteps that by freezing the original weights and learning two small low-rank matrices that represent how each domain perturbs the weight space. These adapters are like compact "patches" that can be swapped in and out as needed. Each patch modifies only a small subspace, so the shared model remains coherent across tasks. Why waste billions of gradient updates rewriting knowledge that already exists when you can preserve it and only store what changes?

Consider a multilingual model trained with LoRA adapters for English, Japanese, and Spanish summarization. Each adapter captures the local linguistic transformations required for that language while leaving the backbone untouched. When the model switches from summarizing Japanese news to English scientific papers, it simply loads a new adapter instead of retraining. It's like changing lenses on a single high-quality camera rather than rebuilding the whole optical system. This modularity is what allows LoRA-equipped models to support hundreds of tasks simultaneously without destructive interference: a feat nearly impossible with pure SFT.

Q2.

The scalability of LoRA depends on the geometry of the pretrained model's manifold. It scales extraordinarily well in efficiency, since each adapter holds only a few million parameters, but it cannot escape the space the base model already understands. LoRA lets us train thousands of domain-specific models on common GPUs or even laptops, creating the first wave of locally fine-tuned LLMs for enterprises and research labs. The analogy often used in the community is that LoRA is like learning how to play new songs on the same instrument: you can master countless tunes, but you cannot change the instrument's tuning without rebuilding it. So while LoRA scales in number of adaptations, it does not scale in expressivity beyond the backbone's structure.

A striking example came from Stability AI's early LoRA experiments, where artists shared visual-style adapters that could be mixed like painteach one light, cheap, and instantly loadable. The community quickly realized that mixing many LoRAs could produce creative hybrids, yet beyond a few, the results became unstable. This mirrored the deeper truth: low-rank adaptations are composable only up to the curvature of the original model's latent space. The method unlocked personalization and domain specialization at scale, but not full reinvention.