LoRA-PEFT vs Unsloth ile Fine-Tuning Karşılaştırma Raporu

Bu raporda, 'TURKCELL/Turkcell-LLM-7b-v1' modeli üzerinde gerçekleştirilen iki farklı fine-tuning yöntemi karşılaştırılmıştır:

1. PEFT kütüphanesi ile LoRA (Low-Rank Adaptation) yöntemi kullanılarak yapılan parameter-efficient fine-tuning.
2. Unsloth kütüphanesi ile yapılan hafifletilmiş ve optimize edilmiş fine-tuning.

Her iki yöntemle eğitilen modeller daha sonra Q4 (4-bit quantization) formatına çevrilmiş ve GGUF formatına dönüştürülerek Ollama ortamında çalıştırılmıştır.

Ortak Noktalar

Özellik	Açıklama
Temel Model	TURKCELL/Turkcell-LLM-7b-v1
Veri Seti	300 özel Türkçe örnek
Son Format	GGUF (Q4 Quantized)
Kullanım Ortamı	Ollama (yerel inference motoru)

1. PEFT + LoRA ile Fine-Tuning

Avantajları

- Hafif ve kaynak dostudur, düşük VRAM ile çalışabilir.
- Hızlı eğitilebilir; kısa sürede sonuç alınabilir.
- Mevcut modelin yalnızca küçük kısmı eğitildiği için daha stabil çalışır.
- LoRA adaptörleri ayrı olarak saklanabilir ve kolayca entegre edilebilir.

Dezavantajları

- GGUF için LoRA ağırlıklarının temel modele birleştirilmesi gerekir.
- Bazı durumlarda tam performans alınamayabilir.
- Dönüştürme sırasında format uyuşmazlığı veya dosya kaynaştırma zorlukları yaşanabilir.

2. Unsloth ile Fine-Tuning

Avantajları

- Eğitim süreci çok optimize edilmiştir, özellikle Colab gibi ortamlarda hızlı sonuç verir.
- LoRA + PEFT altyapısını içerdiği halde tek paket içinde her şeyi otomatik yapar.

- Q4 quantization ile doğrudan entegre çalışabilir; ek birleşim gerekmez.
- Eğitim sonunda doğrudan GGUF'e dönüştürmeye daha elverişlidir.

Dezavantajları

- Şu an için bazı modellerle sınırlı destek olabilir.
- Geliştirme döngüsü PEFT kadar topluluk desteğine sahip değildir.
- Özelleştirme seçenekleri sınırlı olabilir (örneğin özel PEFT config dosyası).

GGUF Formatı ve Ollama Üzerinde Kullanım

Her iki yöntemle eğitilen modeller GGUF formatına başarıyla çevrilmiş ve Ollama üzerinde test edilmiştir.
- LoRA-PEFT yöntemiyle eğitilen modelde GGUF dönüşümü öncesi adapter ağırlıkları ana modele birleştirilmiştir.
- Unsloth yöntemi, model çıktısını doğrudan GGUF uyumlu olarak optimize ettiğinden süreç daha kısa ve kolay olmuştur.

Sonuç ve Öneriler

Eğer hızlı prototipleme, düşük donanım kullanımı ve otomatik süreç istiyorsanız **Unsloth** yöntemi çok daha uygundur. Ancak esnek yapılandırma, topluluk desteği ve geniş model desteği sizin için önemliyse **PEFT + LoRA** yaklaşımı önerilir.

Her iki yöntem de küçük veri setleri için yeterli performans sağlamış ve GGUF formatına başarılı şekilde aktarılmıştır.

Riskler ;

Fine-tuning sürecinde kullanılan veriler sentetik olarak oluşturulmuş ve gerçek dünya örneklerinden sınırlı şekilde beslenmiştir. Bu durum, modelin bağlamı doğru kavrama yeteneğini kısıtlamış ve genel performansı üzerinde olumsuz etki yaratmıştır. Ayrıca, donanım ve kaynak kısıtları nedeniyle model 4-bit quantization (Q4) ile sıkıştırılarak GGUF formatına dönüştürülmüştür. Bu işlem inference hızını artırsa da, modelin hassasiyetini ve doğruluğunu bir miktar düşürmüştür. Az sayıda ve yapay örneklerle eğitilen modelin gerçek dünya senaryolarındaki genelleme kabiliyeti sınırlı kalmıştır. Bu durum, özellikle karmaşık veya düşük frekansta karşılaşılan talepler karşısında modelin yetersiz cevaplar üretmesine neden olabilir.