# Hugging Face Transformers, PEFT, LoRA, và QLoRA

## Hugging Face Transformers

**Hugging Face Transformers** là một thư viện mã nguồn mở nổi tiếng cung cấp các mô hình ngôn ngữ hiện đại (LLMs) đã được huấn luyện sẵn như BERT, GPT, T5, RoBERTa, BLOOM, v.v.
Thư viện này hỗ trợ nhiều tác vụ NLP như: phân loại văn bản, sinh văn bản, dịch, hỏi đáp, v.v.

### Ưu điểm:

- Dễ sử dụng với API thống nhất.
- Có kho mô hình (model hub) phong phú trên https://huggingface.co/models.
- Tương thích với PyTorch, TensorFlow, JAX.

## PEFT (Parameter-Efficient Fine-Tuning)

**PEFT** là viết tắt của *Parameter-Efficient Fine-Tuning*, tức là kỹ thuật **tinh chỉnh mô hình một cách tiết kiệm tham số**.
Thay vì tinh chỉnh **toàn bộ** mô hình (gồm hàng trăm triệu đến hàng tỷ tham số), PEFT chỉ cập nhật **một phần nhỏ**, giúp:

- Tiết kiệm bộ nhớ.
- Nhanh hơn khi huấn luyện.
- Dễ dàng áp dụng với các mô hình lớn.

PEFT phổ biến trong các trường hợp bạn muốn cá nhân hóa mô hình hoặc áp dụng mô hình vào một domain cụ thể mà không cần tốn quá nhiều tài nguyên.

## LoRA (Low-Rank Adaptation)

**LoRA** là một kỹ thuật cụ thể trong PEFT, được dùng để **thêm các ma trận học nhỏ (low-rank matrices)** vào một số lớp của mô hình.
Thay vì cập nhật toàn bộ ma trận trọng số, LoRA chỉ học một phần nhỏ thay thế.

### Cách hoạt động:

- Chèn thêm hai ma trận A và B nhỏ (low-rank) vào trong quá trình huấn luyện.
- Trọng số gốc không thay đổi, chỉ có A và B được học.

- Khi suy luận (inference), các ma trận này sẽ kết hợp lại để mô phỏng hành vi đã tinh chỉnh.

### Ưu điểm:

- Ít tham số cần huấn luyện (ví dụ chỉ 1-2% so với full fine-tuning).
- Có thể chia sẻ hoặc swap các phần LoRA như plugin.

## QLoRA (Quantized LoRA)

**QLoRA** là sự kết hợp giữa:

- **LoRA** (để tinh chỉnh một phần nhỏ).
- **Quantization (lượng tử hóa)** – giảm số bit dùng để lưu trọng số mô hình (ví dụ từ float32 xuống int4).

QLoRA cho phép bạn:

- Tinh chỉnh các mô hình **rất lớn** (7B, 13B, 65B tham số) **trên GPU thương mại như 1 x 24GB VRAM**.
- Sử dụng ít RAM, ít GPU.
- Đạt hiệu suất gần như tinh chỉnh đầy đủ.

QLoRA đã được dùng trong nhiều mô hình hiệu suất cao như Guanaco, RedPajama, v.v.

## Tổng kết:

| Thuật ngữ | Ý nghĩa | Lợi ích chính |
|---|---|---|
| Transformers | Thư viện mô hình NLP mạnh mẽ | Dễ sử dụng, nhiều mô hình sẵn |
| PEFT | Tinh chỉnh tiết kiệm tham số | Nhanh, tiết kiệm tài nguyên |
| LoRA | Cách tinh chỉnh trong PEFT | Chỉ học ma trận nhỏ, hiệu quả |
| QLoRA | LoRA + lượng tử hóa mô hình | Tinh chỉnh mô hình lớn trên máy nhỏ |

---

Phiên bản #1
Được tạo 7 tháng 5 2025 10:14:27 bởi Đỗ Ngọc Tú
Được cập nhật 7 tháng 5 2025 10:16:15 bởi Đỗ Ngọc Tú