Guida ai Parametri LoRA (Versione Aggiornata)

Cos'è LoRA?

LoRA (Low-Rank Adaptation) è una tecnica di fine-tuning efficiente che riduce il numero di parametri trainabili inserendo matrici di rango basso nelle trasformazioni lineari del modello originale. Invece di aggiornare tutti i pesi del modello, LoRA aggiorna solo queste matrici di rango basso, riducendo significativamente la memoria richiesta e accelerando il training.

Parametri Principali

Rank (r)

- **Descrizione**: Determina il rango delle matrici di adattamento LoRA
- **Valore Attuale**: 16
- **Impatto**:
  - Aumenta la capacità di apprendimento
  - Incrementa leggermente la memoria richiesta
  - Migliora la qualità delle didascalie

Alpha

- **Descrizione**: Controlla l'impatto dell'adattamento LoRA
- **Valore Attuale**: 32
- **Impatto**:
  - Aumenta l'effetto dell'adattamento
  - Bilancia il rank per una migliore stabilità
  - Ottimizzato per il task di captioning

Target Modules

- **Descrizione**: Moduli del modello a cui viene applicato LoRA
- **Valore Attuale**: ["q_proj", "v_proj", "k_proj", "o_proj"]
- **Impatto**:
  - Adattamento completo dei componenti di attenzione
  - Migliore capacità di apprendimento
  - Stabilità del training

Dropout

- **Descrizione**: Tasso di dropout applicato durante il training LoRA
- **Valore Attuale**: 0.05
- **Impatto**:
  - Prevenzione overfitting
  - Migliore generalizzazione
  - Stabilità del training

Ottimizzazioni Implementate

Quantizzazione

- **Tipo**: 4-bit (nf4)
- **Compute dtype**: bfloat16
- **Double quantization**: abilitata
- **Vantaggi**:
  - Riduzione significativa della memoria
  - Mantenimento della qualità
  - Supporto per modelli grandi

Batch Size e GPU

- **Batch size per GPU**: 4
- **GPU totali**: 2
- **Batch effettivo**: 8
- **Vantaggi**:
  - Ottimizzazione memoria
  - Training stabile
  - Nessun gradient accumulation

Learning Rate

- **Valore**: 1e-5
- **Scheduler**: cosine
- **Warmup ratio**: 0.05
- **Vantaggi**:
  - Convergenza stabile
  - Prevenzione overfitting
  - Migliore generalizzazione

Risultati Attesi

Performance

- **Gemma 2 9B IT**:
  - Loss di validazione: 0.310
  - Convergenza rapida
  - Qualità didascalie elevata
- **Llama 3.1 8B**:
  - Loss di validazione: 0.516
  - Training stabile
  - Buona qualità didascalie

Efficienza

- **Parametri trainabili**: <0.1% del totale
- **Memoria GPU**: ~48GB
- **Training time**: Ottimizzato per multi-GPU

Best Practices

Configurazione

1. Iniziare con rank=16 e alpha=32
2. Applicare LoRA a tutti i componenti di attenzione
3. Utilizzare dropout=0.05
4. Implementare early stopping

Monitoraggio

1. Tracciare loss di training e validazione
2. Monitorare utilizzo memoria
3. Verificare qualità didascalie
4. Controllare stabilità training

Ottimizzazione

1. Aggiustare rank e alpha in base ai risultati
2. Modificare target modules se necessario

3. Ottimizzare batch size per memoria
4. Regolare learning rate se richiesto