Optimization Techniques of Deep Learning Models for Visual Quality Improvement

Tecniche di Ottimizzazione di Reti Neurali per il Miglioramento della Qualità Visuale

Candidate:	Lorenzo Palloni	lorenzo.palloni@stud.unifi.it
Supervisor:	Marco Bertini	marco.bertini@unifi.it
Co-supervisor:	Leonardo Galteri	leonardo.galteri@unifi.it
Co-supervisor:	Donatella Merlini	donatella.merlini@unifi.it

Abstract [English]

This thesis examines the efficacy of quantization techniques for enhancing the inference speed and reducing the memory usage of deep learning models applied to video restoration tasks. The research investigates the implementation and evaluation of post-training quantization using TensorRT, an NVIDIA tool for inference optimization. The results indicate that reducing the precision of weights and activations substantially decreases computational complexity and memory requirements without compromising performance. In particular, the INT8-optimized UNet and SRUNet models achieve 2.38X and 2.26X speedup compared to their plain implementations, respectively, while also achieving memory consumption reductions of 63.3% for UNet and 53.8% for SRUNet. These findings should contribute to the development of more practical and efficient video restoration models for real-world applications.

Abstract [Italian]

Questa tesi esamina l'efficacia delle tecniche di quantizzazione per migliorare la velocità di inferenza e ridurre l'utilizzo della memoria dei modelli di deep learning per il restauro video. La ricerca indaga l'implementazione e la valutazione della quantizzazione post-allentamento utilizzando TensorRT, un tool sviluppato da NVIDIA per l'ottimizzazione in inferenza. I risultati indicano che la riduzione della precisione numerica dei pesi e delle attivazioni riduce notevolmente la complessità computazionale e lo spazio di memoria occupato senza comprometterne le prestazioni. In particolare, i modelli UNet e SRUNet ottimizzati ad una precisione INT8 raggiungono rispettivamente un aumento di velocità del 2.38X e del 2.26X rispetto alle loro implementazioni originali, mentre raggiungono riduzioni del consumo di memoria fino al 63.3% per la UNet e fino al 53.8% per la SRUNet. Questi risultati dovrebbero contribuire allo sviluppo di modelli per il restauro video più efficienti e quindi più pratici per applicazioni reali.