TinyViT – Descrizione del Modello

TinyViT è un modello Vision Transformer compatto e ad alte prestazioni, progettato per operare efficientemente su dispositivi con risorse limitate (es. dispositivi edge o mobili). È stato introdotto nel paper:

"TinyViT: Fast Pretraining Distillation for Small Vision Transformers"
Menglin Jia et al., 2022
https://arxiv.org/abs/2207.10666

Architettura

TinyViT combina le caratteristiche dei Vision Transformer (ViT) con elementi dei modelli CNN:
- Utilizza una struttura ibrida con convoluzioni nelle fasi iniziali e Transformer blocks nei livelli superiori.
- Impiega finestra di attenzione locale (come in Swin Transformer) per ridurre la complessità computazionale.
- Integra positional encoding locali per catturare informazioni spaziali.

Knowledge Distillation

Per migliorare la precisione mantenendo una dimensione ridotta, TinyViT viene addestrato usando knowledge distillation:
- Il modello teacher è un ViT di grandi dimensioni (es. ViT-G/14, Swin-B, o simili), pre-addestrato su dataset estesi (es. ImageNet-21K).
- Durante l'addestramento, TinyViT imita le predizioni (logits) e talvolta le attivazioni intermedie del teacher, apprendendo così rappresentazioni più robuste.

Applicazioni

TinyViT è ideale per:
- Classificazione d'immagini in tempo reale
- Deploy su dispositivi mobili e embedded
- Visione artificiale edge-aware (es. AR/VR, robotica)