Vision Transformer (ViT) for Medical Imaging

Process medical images with patch embedding and positional encoding, capture global relationships through self-attention

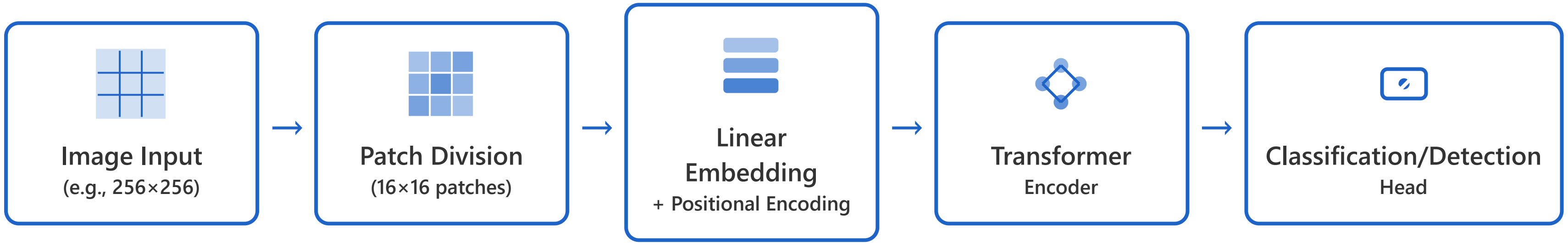

Image Input (e.g., 256×256) → Patch Division (16×16 patches) → Linear Embedding + Positional Encoding → Transformer Encoder → Classification/Detection Head

Self-Attention Mechanism

Learning long-range dependencies across entire image

- Compute relationships between all patches
- Query, Key, Value transformations
- Capture spatial relationships between lesions
- Multi-head attention

Medical Imaging Applications

ViT specialized for medical domain

- High-resolution image processing
- Variable-size input support
- 3D volume extension (3D-ViT)
- Medical-specific pre-training

Advantages

Strengths of ViT

- Global context understanding

Medical ViT Models

Representative medical-specialized ViT

- MedViT: Medical image pre-training

- Scalable architecture
- Large-scale pre-training benefits
- Various downstream tasks

- TransUNet: Segmentation specialized
- Swin Transformer: Hierarchical
- CoTr: CT/MRI reconstruction