Image classification - Model layer

Hugging Face Model Zoo Options

CNN Families (classic baselines)

- **ResNet-50 / ResNet-101** *Checkpoint:* `microsoft/resnet-50` *Why it matters:* Deep residual connections solved vanishing gradients, still strong baselines.

- **EfficientNet (B0–B7)** *Checkpoint:* `google/efficientnet-b0` *Why it matters:* Compound scaling (depth, width, resolution) → strong performance/efficiency trade-off.

Vision Transformers (modern default)

- **ViT-Base / ViT-Large (patch16/224)** *Checkpoint:* `google/vit-base-patch16-224` *Why it matters:* First pure-transformer image classifier; competitive with CNNs when pretrained on large corpora.

- **DeiT (Data-efficient Image Transformer)** *Checkpoint:* `facebook/deit-base-distilled-patch16-224` *Why it matters:* Distillation tricks make transformers viable with less data; faster training.

- **Swin Transformer** *Checkpoint:* `microsoft/swin-base-patch4-window7-224` *Why it matters:* Hierarchical windows + shifting → better locality modeling than vanilla ViT.

Hybrid / Advanced Architectures

- **ConvNeXt** *Checkpoint:* `facebook/convnext-base-224` *Why it matters:* CNN redesigned with transformer-era tricks (layer norm, GELU, large kernels).

- **BEiT (BERT for Images)** *Checkpoint:* `microsoft/beit-base-patch16-224` *Why it matters:* Masked image modeling (like MLM in NLP) → powerful self-supervised pretraining.

- **CLIP (multimodal, classification via zero-shot)** *Checkpoint:* `openai/clip-vit-base-patch32` *Why it matters:* Joint image–text embeddings enable zero-shot classification and flexible labeling.

Architectural Innovations

- **CNNs (ResNet, EfficientNet):** Inductive biases (convolutions, pooling) → data-efficient, fast convergence.

- **ViTs (ViT, DeiT, Swin):** Attention-only, no convolutions; scalable with pretraining, interpretability via attention maps.
- **Hybrid (ConvNeXt):** CNNs reimagined with transformer-era training strategies.
- **Self-supervised Transformers (BEiT, MAE):** Learn representations without labels (masked image modeling).
- **Multimodal (CLIP):** Align vision and text → flexible zero-shot classification.
