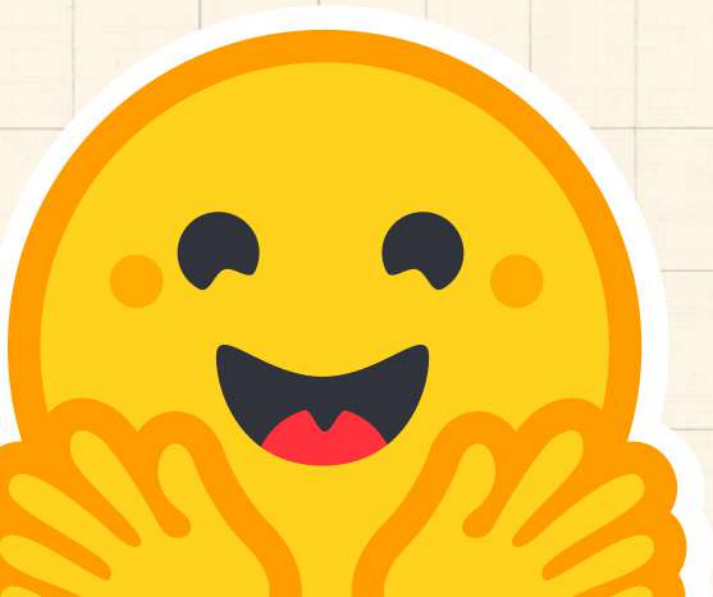
Tomer Biton

HF model of the weekend

vit-base-patch16-224



*google/vit-base-patch16-224

Tomer Biton

What ViT Actually Is

Instead of pixels → CNN → classifier
ViT cuts the image into patches → treats them like tokens → uses self-attention to understand the whole scene.
Why it matters:

- No convolutions
- Scales extremely well
- Powerful general-purpose baseline for most image tasks
- A great way to understand how vision transformers “think"

Tomer Biton

How It Works Internally

1. Resize image → 224 × 224
2. Normalize (mean 0.5, std 0.5)
3. Convert to patches (16×16)
4. Embed each patch
5. Add position embeddings
6. Pass through Transformer layers
7. Read prediction from CLS token

Tomer Biton

What I built

A simple but effective CLI image classifier:

- ✓ Load image from file or URL
- ✓ Process it with ViTImageProcessor
- ✓ Predict the top class from ImageNet-21k
- ✓ Get clean, readable output (e.g., “tabby cat”, “sports car”, “teddy bear”)

Best for:

- **Image classification**
- **Feature extraction**
- **Transfer learning**
- **Serving as a backbone for custom classifiers**
- **Quick prototypes and demos**

This is a beast for simple vision tasks.

Tomer Biton

What I Learned This Weekend

Small model, big takeaways.

- No convolutions → pure transformers
- Generalizes extremely well
- Lightweight for its accuracy
- Perfect for fast experiments
- Great intro to vision transformers

🔥 Definitely one of the best “first vision models” to explore.