AN IMAGE IS WORTH 16X16 WORDS:

TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

v-0.1

Seminário

Rafael Dias da Silva

Orientador: Adriano Maurício de Almeida Cortes

30 de abril de 2025

Motivação

Motivação

- **Problema**: CNNs dominavam visão computacional, mas têm limitações:
  - Indutores de viés local (receptive fields fixos).
  - Dificuldade em modelar dependências de longo alcance.
- **Oportunidade**: Sucesso dos Transformers em NLP (ex.: BERT, GPT).
- **Pergunta-chave**: ”Transformers podem substituir CNNs em visão computacional?”

Arquitetura

- Divisão da imagem em patches de 16×16.
- Projeção linear + positional embeddings.
- Pilha de camadas Transformer (auto-atenção + MLP).

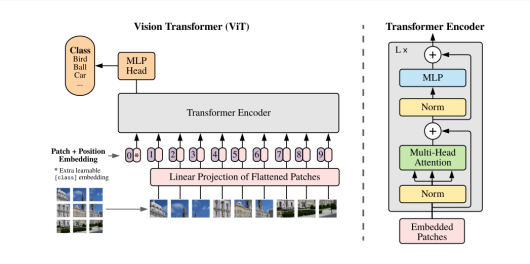


Figura 1: Arquitetura do Vision Transformer. Fonte: Dosovitskiy et al. (2020).

Tokenização

- **Passo 1**: Dividir imagem em N patches não sobrepostos.

$$N = \frac{H \times W}{P^2} \quad \text{(ex.: } 224 \times 224 \rightarrow 196 \text{ patches)}$$

- **Passo 2**: Projeção linear de cada patch para espaço latente:

$$\mathbf{z}_i = \mathbf{E} \cdot \text{Flatten}(\mathbf{x}_i) + \mathbf{p}_i$$

- **Analogia com NLP**: Patches = ”palavras visuais”.

Estratégia de Treinamento

- **Pré-treinamento**: Em datasets massivos (JFT-300M ou ImageNet-21k).
- **Fine-tuning**: Adaptação para datasets menores (ex.: ImageNet).
- **Híbrido**: Opcionalmente, usar feature maps de CNNs como entrada.

Embedding de Posição

- **Problema**: Transformers não têm noção inata de espaço.
- **Solução**: Embeddings posicionais 1D aprendidos.

$$z = [x_{\text{class}}; x_1 E; \ldots; x_N E] + E_{\text{pos}}$$

- **Discussão**: Embeddings 2D não melhoraram resultados.

MULTIHEAD SELF-ATTENTION

- Mecanismo de atenção com k cabeças paralelas:

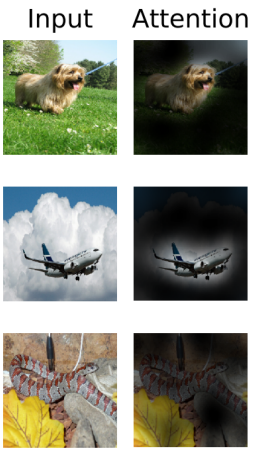
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

- **Vantagem**: Captura múltiplos tipos de relações espaciais.

Experimentos

- **Datasets**: JFT-300M (303M imagens), ImageNet, CIFAR.
- **Resultados-Chave**:
  - ViT-L/16: 87.76% top-1 (ImageNet).
  - 2-4x mais eficiente que CNNs equivalentes.

Mapa de atenção



Estudo de escala

- **Modelos**: ViT-Base, Large, Huge.
- **Conclusão**:
  - Performance escala com tamanho do modelo/dados.
  - ViTs superam CNNs com datasets grandes o suficiente.

Limitações

- Requer datasets muito grandes para pré-treinamento.
- Custo computacional alto para alta resolução.
- Menos eficiente que CNNs em dados limitados.

Auto supervisão

Conclusão

- **Contribuições**:
  - Prova que Transformers puros podem ser SOTA em visão.
  - Abre caminho para arquiteturas unificadas (NLP + visão).
- **Futuro**: Versões hierárquicas (ex.: Swin Transformer).

Referências i


Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020).
An image is worth 16x16 words: Transformers for image recognition at scale.