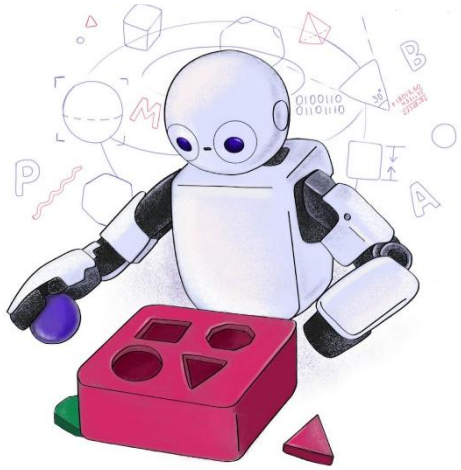
TP558 - Tópicos avançados em Machine Learning: Convolutional Vision Transformer (CvT)



Inatel

LUIS ENRIQUE CARDOZO RAMIREZ

Introdução

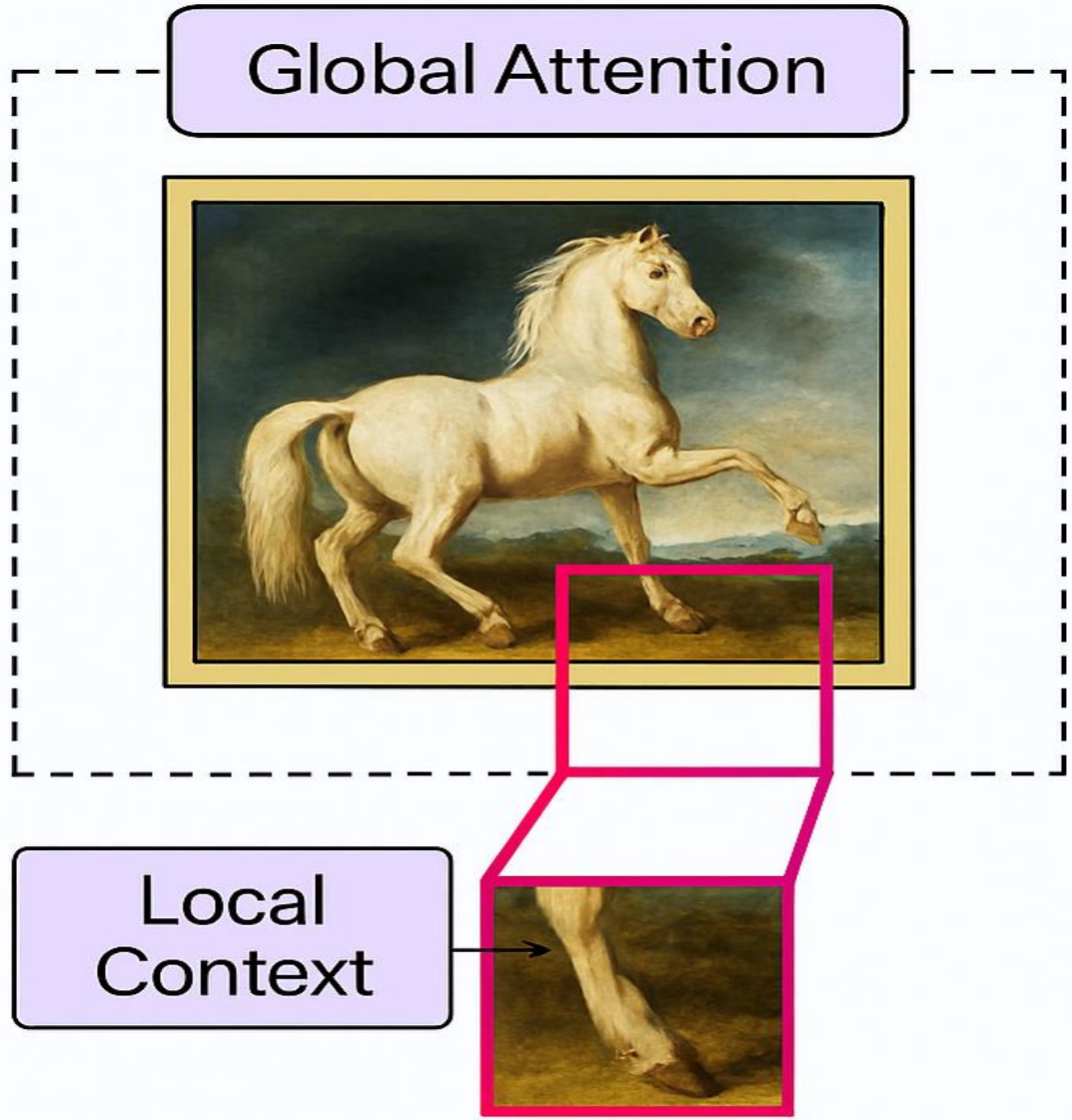
- Já pensaram em como dar a uma inteligência artificial o poder de ver o mundo? Os modelos mais poderosos de visão por computador, como o Vision Transformer (ViT), precisam de uma **quantidade gigantesca de dados para se saírem bem.**
- Mas e se os dados que temos não são suficientes? Como podemos fazer com que essas arquiteturas **entendam a imagem de forma mais "natural",** como nós?

Introdução

A resposta está em uma arquitetura hibrida: o **Convolutional vision Transformer (CvT).**

Em vez de apenas tokens, ele usa convoluções, que são especialistas em capturar os **detalhes locais**, **como as texturas e as bordas de um objeto.**

O resultado é uma IA que, assim como nós, vé o quadro geral (atenção global) e também se atenta aos détalhes (contexto local), tornando-se mais eficiente e poderosa.



Introdução

O Cenário dos Transformers em Visão Computacional

- **Transformers:** Originalmente dominantes em Processamento de Linguagem Natural (PNL).
  - Conhecidos por capturar dependências de longo alcance e contexto global.
- **Vision Transformer (ViT):** O primeiro a aplicar exclusivamente a arquitetura Transformer para classificação de **imagens em larga escala**, alcançando desempenho competitivo.
  - **Como funciona:** Decompõe uma imagem em sequências de "patches" (tokens) não sobrepostos de tamanho fixo, que são então processados por **camadas Transformer com módulos de autoatenção multi-cabeça** (MHSA) e redes feed-forward (FFN).
  - **Atenção Multi-Cabeça (MSA):** Permite que cada patch "olhe" para todos os outros, entendendo relações em toda a imagem e capturando diferentes perspectivas simultaneamente.

Introdução

O Que São CNNs

- Redes Convolucionais (CNNs): Arquiteturas dominantes em visão por computador que se destacam na captura de estruturas locais.
- Propriedades-chave: Forçam a captura de estruturas locais usando campos receptivos locais, **pesos compartilhados e subamostragem espacial**, o que resulta em invariância a translação, escala e distorção.
- Hierarquia: Aprendem padrões visuais hierarquicamente, de arestas simples a padrões semânticos complexos.
- O problema do ViT: Quando treinados com menos dados, seu desempenho é inferior ao de CNNs de tamanho similar, pois lhes faltam as propriedades desejáveis das CNNs.

Introdução

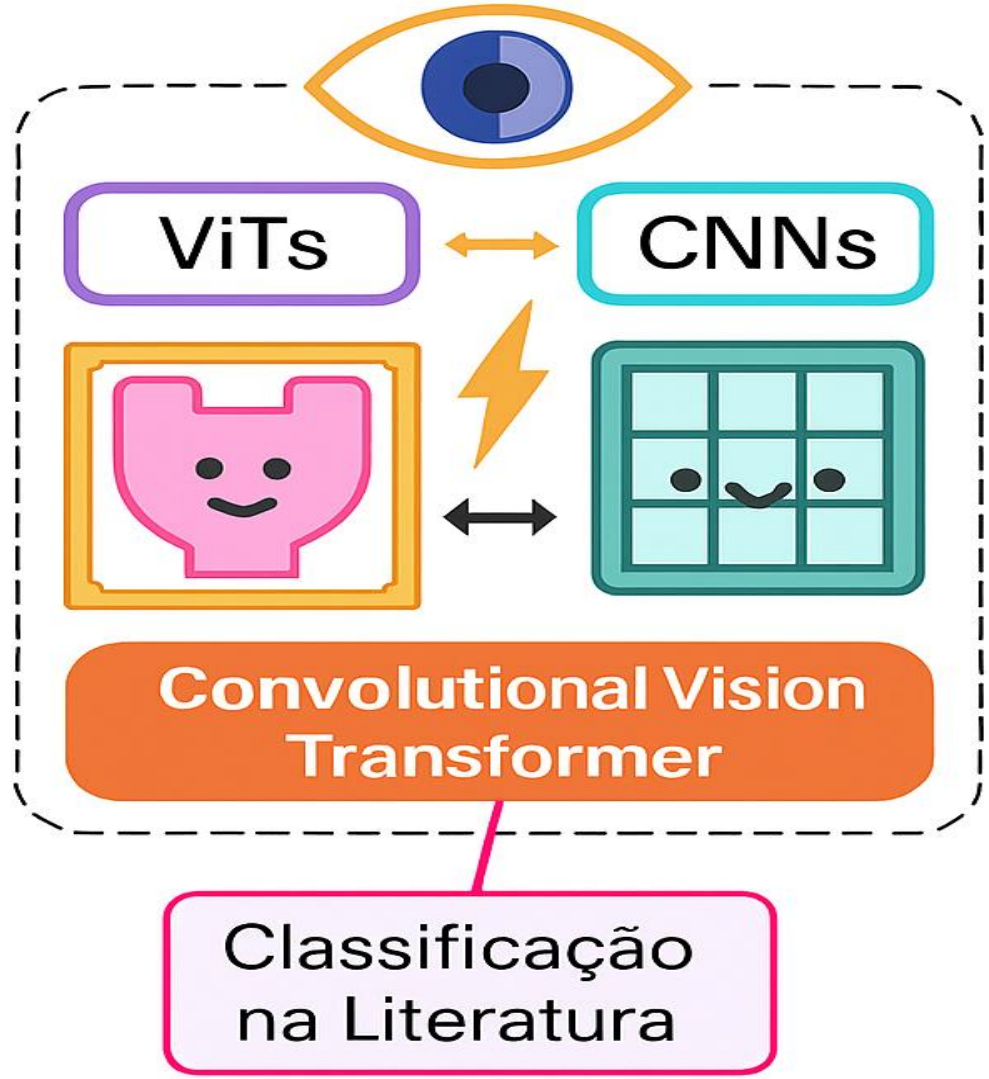
O Desafio dos ViTs Puros e as Vantagens das CNNs

Introdução

A Solução: O Que é o **Convolutional Vision Transformer (CvT)?**

Definição: O Convolutional Vision Transformer (CvT) é uma nova arquitetura de deep learning que combina estratégicamente as **forças das Redes NeuraiscConvolucionais (CNNs)** e dos **Vision Transformers (ViTs).**

Objetivo Principal: Superar as limitações dos ViTs puros, introduzi-ndo convoluções para incorporar o biás indutivo local das CNNs, enquanto mantém os Transformers, como atenção dinâmica, contexto global e melhor generalização.



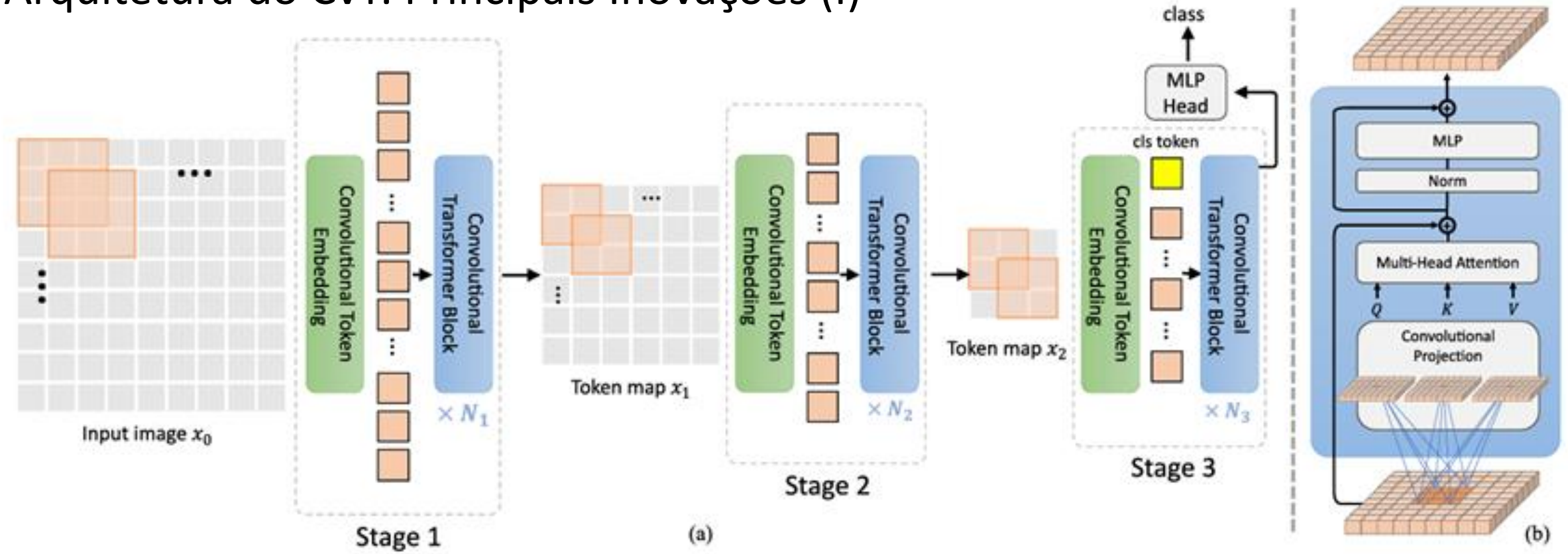
Arquitetura e funcionamento

Arquitetura do CvT: Principais Inovações (I)

- **1. Hierarquia de Transformers com Embedding Convolucional de Tokens**
  - **Função:** Substitui o método de particionamento de patches não sobrepostos do ViT.
  - **Mecanismo:** A imagem de entrada ou os mapas de tokens 2D de estágios anteriores são submetidos a uma camada de convolução sobreposta (com stride).
  - **Benefícios:**
    - **Captura de Informações Locais:** Essencial para modelar contextos espaciais, desde bordas de baixo nível até primitivas semânticas de ordem superior, em uma abordagem hierárquica.
    - **Downsampling Eficiente:** Reduz progressivamente o comprimento da sequência de tokens e aumenta a dimensão das características entre os estágios, semelhante ao que ocorre nas CNNs.
    - **Flexibilidade:** Permite ajustar a dimensão das características e o número de tokens em cada estágio, variando os parâmetros da operação de convolução.

Arquitetura e funcionamento

Arquitetura do CvT: Principais Inovações (I)



A Figura 1 mostra o pipeline geral, enfatizando a estrutura hierárquica em três estágios. Cada estágio utiliza uma camada de Embedding Convolucional de Tokens para transformar a imagem de entrada ou mapas de tokens anteriores em representações espaciais mais ricas, reduzindo a resolução espacial (via stride) enquanto aumenta a dimensionalidade das características.

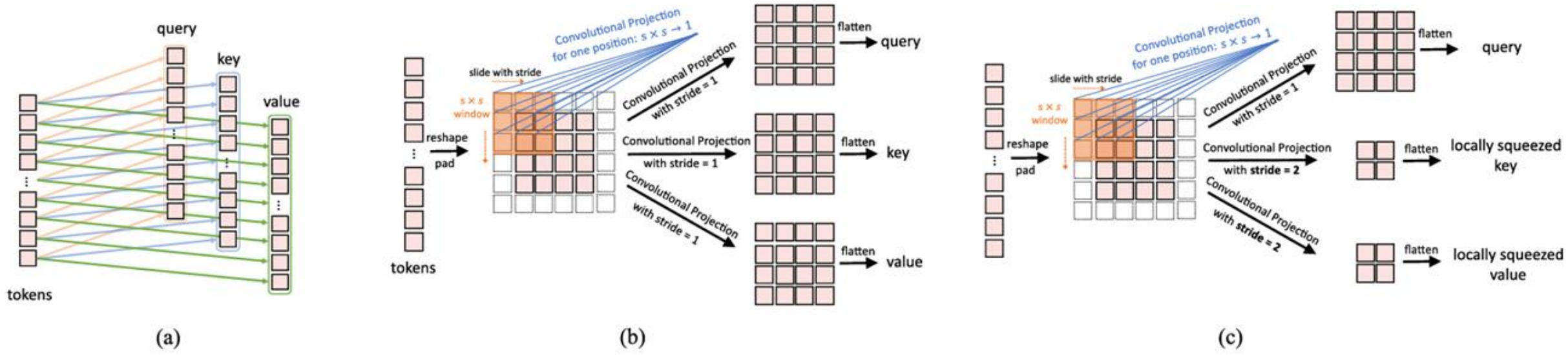
Arquitetura e funcionamento

Arquitetura do CvT: Principais Inovações (II)

- **2. Bloco do Transformer Convolucional com Projeção Convolucional**
- **Função:** A projeção linear padrão, utilizada antes de cada bloco de autoatenção em um Transformer para Q (query), K (key) e V (value), é substituída por uma **projeção convolucional**.
- **Mecanismo:** Emprega uma convolução separável em profundidade (***depth-wise separable convolution***) de tamanho s x s em um mapa de tokens 2D.
- **Benefícios de Eficiência:**
  - **Contexto Local Aprimorado:** Captura ainda mais contexto espacial local e reduz a ambiguidade semântica no mecanismo de atenção.
  - **Redução de Custo Computacional:** **Permite a subamostragem das matrizes K e V utilizando um stride maior que 1** (por exemplo, stride de 2, reduzindo 4x o custo). Isso resulta em uma melhoria de eficiência de 4x ou mais na operação MHSA, com degradação mínima de desempenho.
  - As convoluções compensam a perda de informação causada pela redução de resolução.

Arquitetura e funcionamento

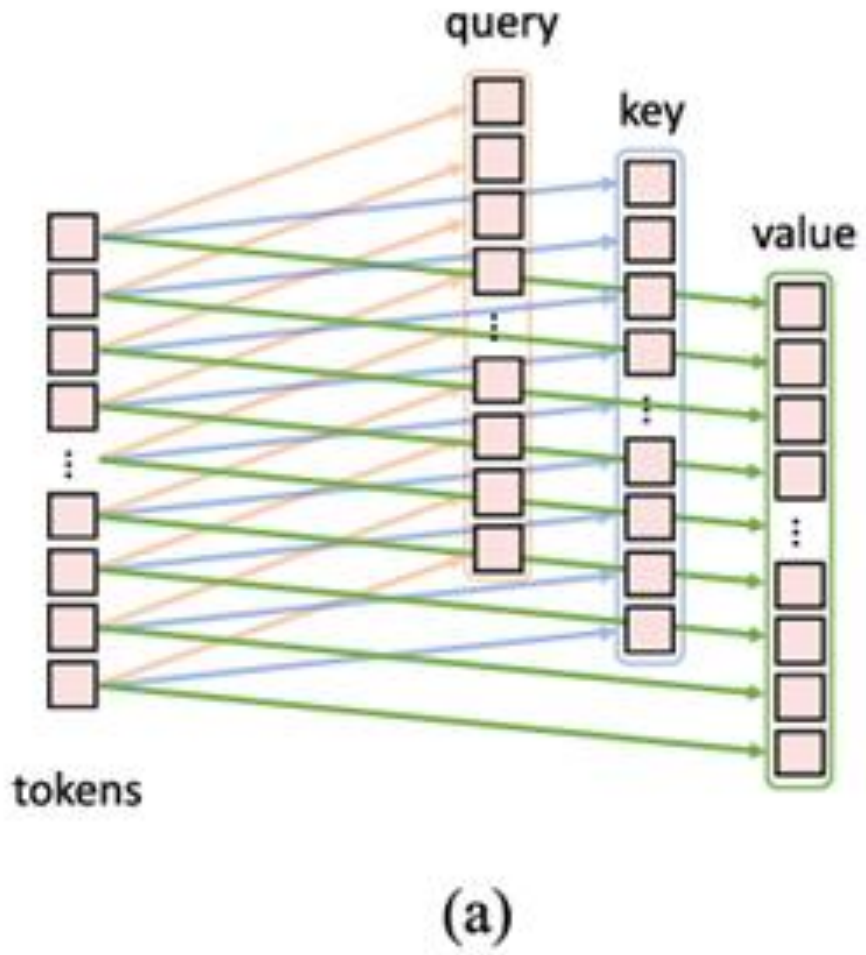
Arquitetura do CvT: Principais Inovações (II)



A Figura 2 ilustra as três abordagens de projeção usadas no mecanismo de atenção do Transformer

Arquitetura e funcionamento

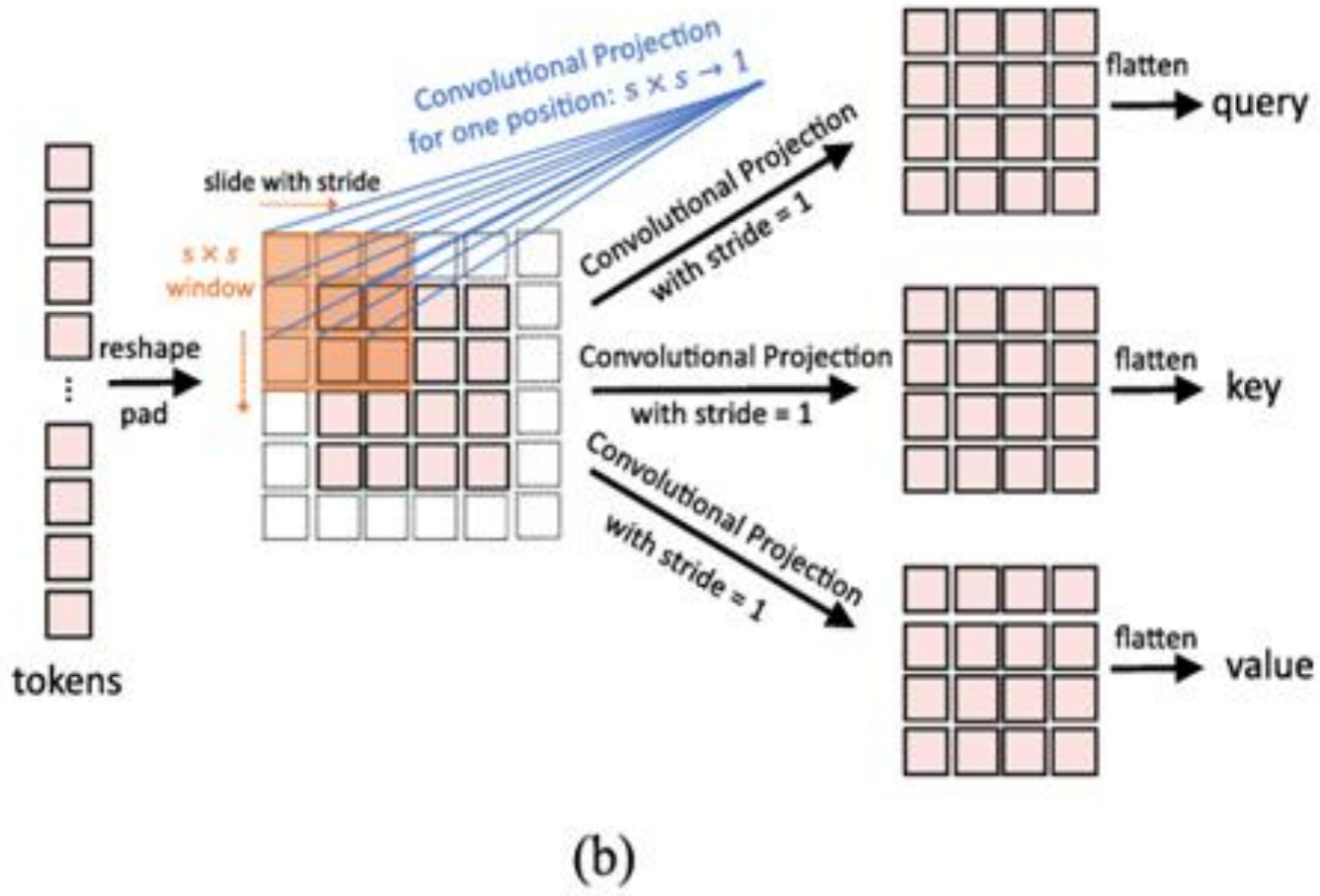
Arquitetura do CvT: Principais Inovações (II)



A parte (a) mostra a projeção linear do ViT original, onde os tokens são projetados diretamente em Q, K e V sem considerar a estrutura espacial local, o que limita a capacidade de capturar correlações locais.

Arquitetura e funcionamento

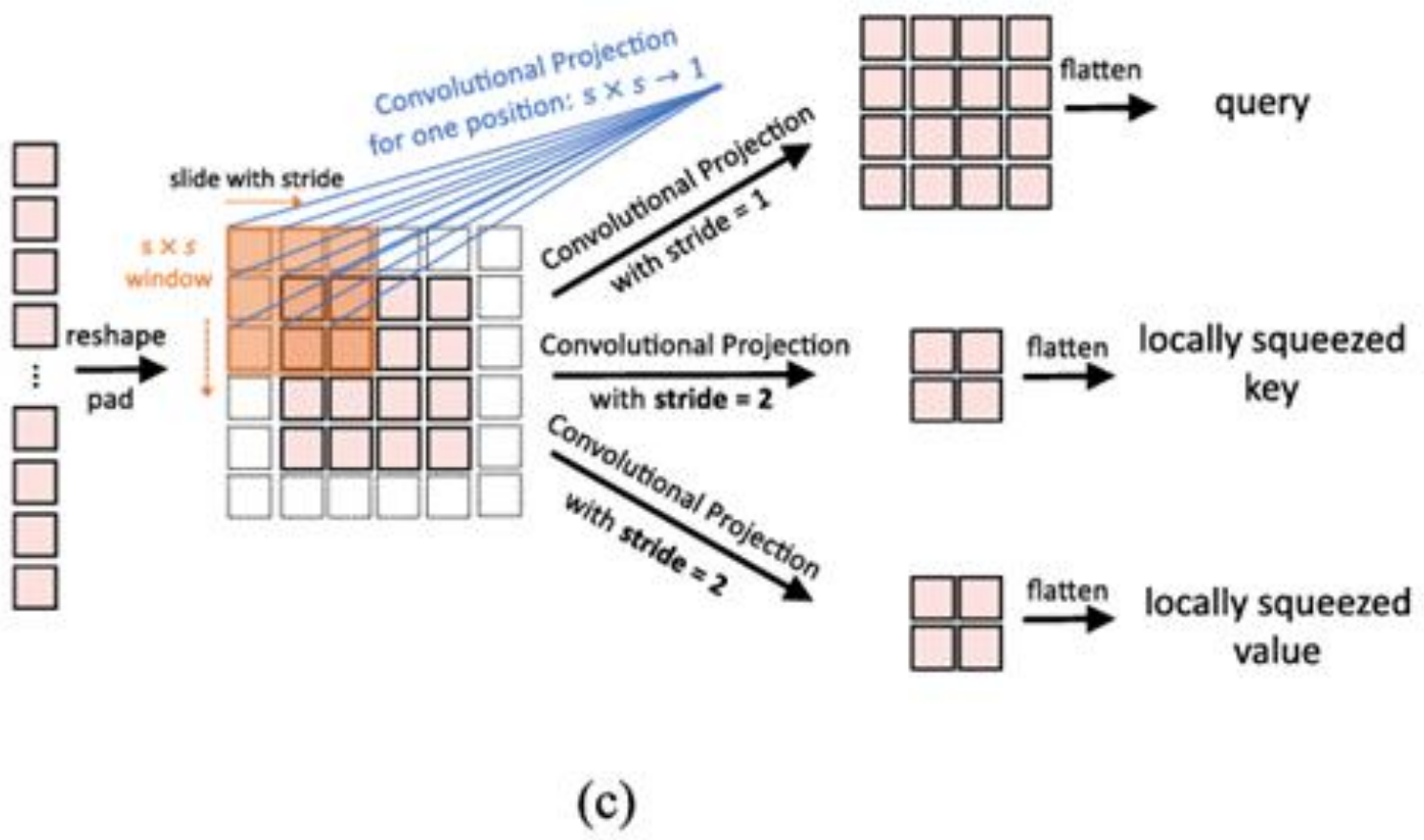
Arquitetura do CvT: Principais Inovações (II)



A parte (b) introduz a projeção convolucional, onde os tokens são reshapeados em um mapa 2D, convoluídos com uma janela `s x s` e stride 1, e achatados, permitindo que o modelo incorpore informações espaciais locais antes da atenção.

Arquitetura e funcionamento

Arquitetura do CvT: Principais Inovações (II)



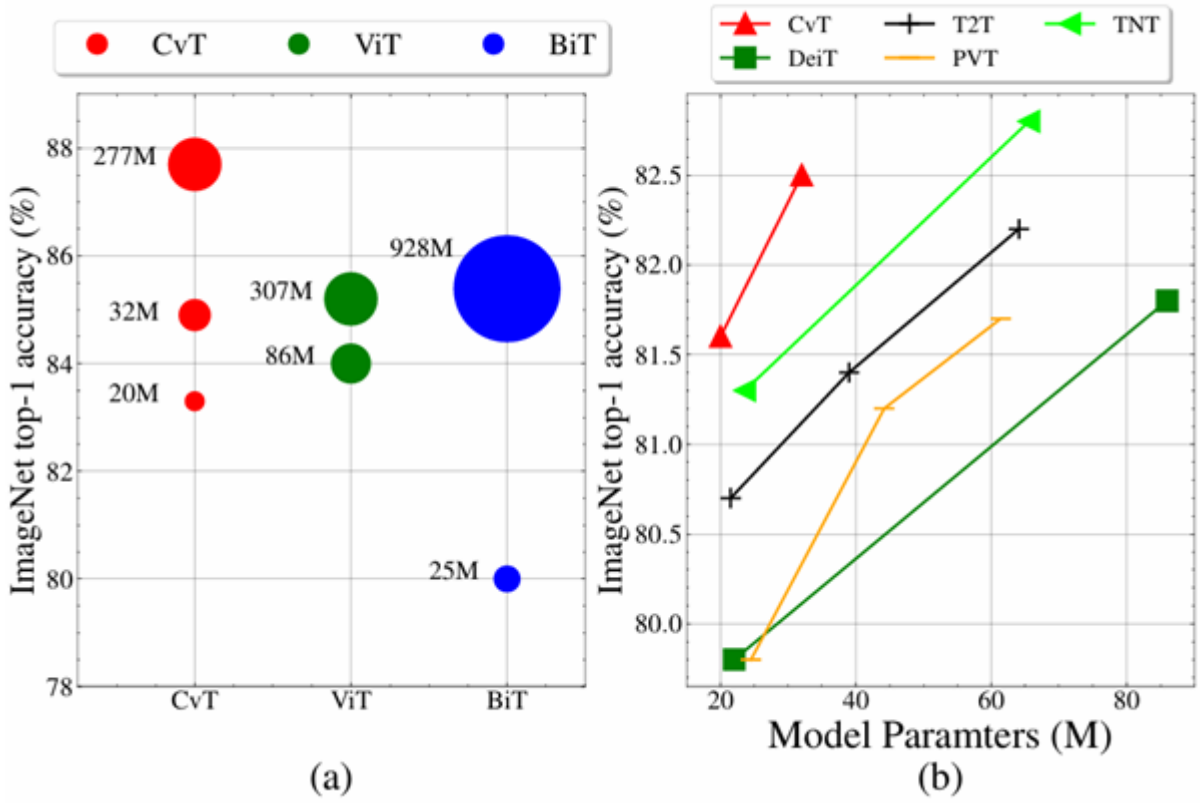
(c)

A parte (c), a projeção convolucional espremida (padrão no CvT), adiciona um stride maior, após a convolução, reduzindo a resolução de K e V e otimizando a eficiência computacional ao manter a captura de contexto local. Essa escolha padrão reflete o equilíbrio entre desempenho e eficiência no CvT.

RESULTADOS

- **Vantagens e Desempenho**
- **Desempenho de Ponta e Eficiência:** O CvT atinge um desempenho superior aos Vision Transformers e ResNets existentes na classificação ImageNet-1k, utilizando menos parâmetros e FLOPs.
- **Invariância Reforçada:** Herda propriedades desejáveis das CNNs, como invariância a deslocamento, escala e distorção.
- **Remoção de Positional Embeddings:** Uma vantagem significativa é que o CvT **não requer embeddings posicionais explícitos**.
  - As convoluções incorporadas permitem que a rede modele relações espaciais localmente.
  - Simplifica o design e facilita a adaptação para tarefas de visão com resoluções de entrada variáveis, sem a necessidade de re-projeto do embedding.
- **Estrutura Hierárquica e Representação Rica:** O design multiestágio com convoluções permite downsampling espacial e o aprendizado de padrões visuais mais complexos em áreas espaciais maiores, similar às camadas de características das CNNs.
- **Convergência Rápida:** Demonstra uma convergência significativamente mais rápida durante o treinamento em comparação com modelos Transformer originais.

RESULTADOS



- Acurácia Top-1 no ImageNet validação comparada a outros métodos em relação aos parâmetros do modelo.

- A parte (a) compara o CvT com o modelo baseado em CNN BiT e o Transformer-based ViT, ambos pré-treinados no ImageNet-22k. O CvT com 277M de parâmetros alcança 87.7\% de acurácia Top-1, superando ViT (928M) e BiT (307M), demonstrando superioridade em eficiência e desempenho.
- A parte (b) foca em modelos pré-treinados no ImageNet-1k, incluindo DeiT, T2T, PVT e TNT, onde o CvT com até 32M de parâmetros atinge 82.5\% de acurácia, destacando-se pela menor complexidade computacional.

Comparativo: CvT vs. Outras Arquiteturas Híbridas

Comparativo Essencial

Tabela 1: Comparativo: CNNs vs. ViT vs. CvT

Aspecto	**CNN**	**ViT**	**CvT**
Viés Indutivo Local	Forte	Fraco	Forte (via convoluções)
Estrutura Hierárquica	Sim	Não	Sim
Positional Encoding	Não	Requerido	Não Requerido
Mecanismo de Atenção	Nenhum	Autoatenção Global	Autoatenção Convolucional
Eficiência em Alta Res.	Alta	Baixa	Alta

RESULTADOS

Outros ViTs Híbridos (Exemplos)

•**DETR (Detection Transformer):** Utiliza uma CNN como backbone para extrair mapas de características antes de alimentar um encoder-decoder Transformer para detecção de objetos.

•**PVT (Pyramid Vision Transformer):** Incorpora uma estrutura piramidal multiestágio, semelhante às CNNs, para lidar com mapas de características de alta resolução, usando um módulo de *spatial-reduction attention* (SRA).

•**Conformer:** Combina ramos separados de CNN (para percepção local) e Transformer (para características globais), com conexões entre eles.

•**MaxViT (Multi-Axis Attention-based Vision Transformer):** Adota uma estrutura hierárquica com blocos híbridos que consistem em convoluções MBConv e atenção multi-eixo (local e global).

•**CeiT (Convolution-enhanced Image Transformer):** Modifica o esquema de extração de patches, a camada MLP e adiciona uma camada convolucional final para melhor desempenho.

Desempenho Empírico do CvT

Resultados em Imagem e Transfer Learning

Classificação de Imagens (ImageNet-1k):

Tabela 2: Modelos pré-treinados no ImageNet-1k. repositório GitHub.

Modelo	Resolução	Parâmetros (M)	GFLOPs	Top-1 (%)
CvT-13	224x224	20	4.5	81.6
CvT-21	224x224	32	7.1	82.5
CvT-13	384x384	20	16.3	83.0
CvT-32	384x384	32	24.9	83.3

Implementação e Ferramentas

A implementação oficial do CvT está disponível no GitHub. Também pode ser utilizada através da biblioteca Hugging Face transformers.

- **Principais Blocos de Código (Conforme Implementação Oficial):**
  - `_build_projection`: Implementa a **Projeção Convolucional** para as matrizes Query, Key e Value, utilizando convoluções separáveis em profundidade.
  - `ConvEmbed`: Responsável pelo **Embedding Convolucional de Tokens**, processando patches sobrepostos da imagem de entrada com camadas convolucionais.
  - `VisionTransformer`: O bloco central do Transformer, adaptado para usar o `ConvEmbed` para patch embedding e as projeções convolucionais dentro do mecanismo de auto-atenção.
  - `ConvolutionalVisionTransformer`: A arquitetura CvT completa, que compõe múltiplas `VisionTransformer` (estágios) em uma estrutura hierárquica.

Exemplo(s) de aplicação

- Aplicações em Visão Computacional

•**Backbone Versátil:** O CvT serve como um backbone robusto e eficiente para uma ampla gama de tarefas de visão computacional.

•**Classificação de Imagens:** Sua aplicação primária e onde demonstrou desempenho de ponta (ex: ImageNet-1k, ImageNet-22k).

•**Detecção de Objetos e Segmentação de Instâncias:** As modificações introduzidas pelo CvT o tornam adequado para tarefas de previsão densa. Exemplos de outros HVTs já mostraram sucesso em Mask R-CNN e Cascade R-CNN.

Exemplo(s) de aplicação

- Outras Aplicações de HVTs (tendências gerais):

•**Reconhecimento de Imagens/Vídeos:** HVTs são usados para capturar informações globais e locais, melhorando o reconhecimento em cenários complexos.

•**Reconstrução de Imagens:** Em tarefas como super-resolução e *denoising*, combinam a modelagem local das convoluções com a global da autoatenção.

•**Extração de Características:** Para identificar e extrair informações visuais relevantes em diversas aplicações.

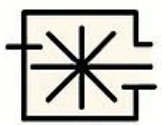
•**Análise de Imagens Médicas:** HVTs mostram grande promessa na segmentação e análise de imagens médicas, onde a captura de detalhes locais e contexto global é crucial.

•**Classificação de Morfologia Galáctica:** Aplicações recentes em astronomia, como classificação de morfologia de galáxias usando CvT.

Vantagens e desvantagens

Vantagens

Desempenho Superior e Eficiéncíã
CvT melhora VIT em permanença e énein ao state-of-the-art em pérmanença e ficiénua. alcançado ote-t ọ imageNer-1k e alcaando et ém pré-desempenno enrintre CNNs e VITs



Remocão de Positional Embeddings
Elimina a necesssidade de emberdádings posicionais explicitos sem pérda de desernpenho, simplificação em esconsanho como

Desvantagens

Complexidade de Implementação
A arquitura hibrida combina CNNs e Transformers, o que pode aumentar a complexidade de implementação e otimização ém comparaçao ao modelos puros,

Custo Computacional Antor mais eficiénte que VITs, ainda herda aguma compléxidade quadridiça da da atenção, podendo ser desafiador e embléntes com recursos limitados ou imagens de alta resolu-

Reusabilidade e Tuning C ébibirbrio entre componentes convolucienais pode comfortar á uṣar toning culdadoso, e a reuitliziiçao de modélos ssimples

Validação e Refutação

O **Convolutional Vision Transformer (CvT)** representa um avanço significativo no campo da visão computacional, unindo com sucesso as melhores características das Redes Neurais Convolucionais (CNNs) e dos Vision Transformers (ViTs).

Através do **Embedding Convolucional de Tokens** e da Projeção Convolucional, o CvT incorpora o viés indutivo local das CNNs, mantendo a capacidade de modelagem de contexto global dos Transformers.

Isso resulta em **desempenho superior e maior eficiência**, com menos parâmetros e FLOPs, e a notável capacidade de dispensar positional embeddings, simplificando o design para tarefas de visão de resolução variável.

O CvT é um **backbone versátil** que pavimenta o caminho para o desenvolvimento de modelos de visão computacional mais robustos, eficientes e adaptáveis a um espectro ainda maior de aplicações práticas.

QUIZ

QUIZ SEMINARIO

GITHUB

Referências

[1] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[2] Li Yuan, Qibin Hou, Zeming Li, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[3] Yanghao Tu, Zhicheng Cai, Zhibin Wang, Yichen Liu, and Dahua Lin. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*, pages 3–20. Springer, 2022.

[4] Li Yuan, Shi Pu, Francis EH Tay, and Jiashi Feng. Ceit: Convolution-enhanced image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 101–110, 2022.

[5] Yixuan Chen, Zhen-Ya Sun, Xu-Kang Zhang, and Yi-Liang Liu. Galaxy morphology classification based on convolutional vision transformer. *Astronomy & Astrophysics*, 2024.

Obrigado!