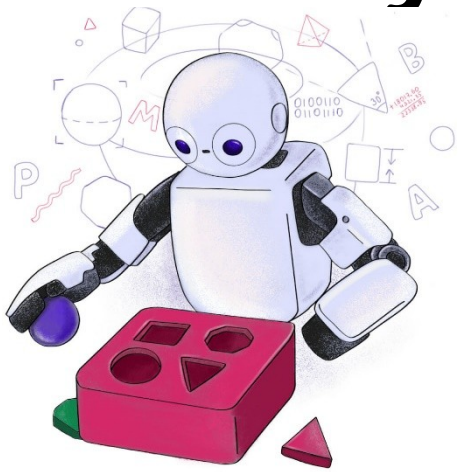
TP558- Tópicos avançados em Machine Learning: *Vision Transformer*



Inatel

Steffie Gabriella Jean Gilles
steffie@mtel.inatel.com

Introdução

- Devido às dificuldades das redes neurais convolucionais em interpretar conjuntos de dados grandes e complicados, foi desenvolvida uma nova arquitetura alternativa, como os **transformadores de visão** ViT.
- Essa arquitetura foi inicialmente projetada e desenvolvida para o processamento de linguagem natural (NPL), mas agora foi modificada para aplicações vinculadas ao processamento de imagens por meio de ViTs.

O que propõe a resolver o algoritmo de ViT?

- O algoritmo de **Vision Transformer (ViT)** propõe resolver problemas de visão computacional, especificamente em tarefas de classificação de imagens.

- O algoritmo do **Vision Transformer (ViT)** propõe resolver o problema de reconhecimento de imagem sem a necessidade de arquiteturas híbridas que combinam convoluções e mecanismos de atenção.

O que propõe a resolver o algoritmo de ViT?

A contextualização do problema que o ViT se propõe resolver é o desafio de reconhecimento de imagem de forma mais eficiente, escalável e livre de viéses arquiteturais, demonstrando que um modelo baseado em Transformers pode alcançar resultados competitivos em tarefas de visão computacional, mesmo sem a incorporação de inductive biases específicos de imagens além do passo inicial de extração de patches.

Fundamentação teórica

Conceitos teóricos fundamentais por trás do ViT

- Transformers e mecanismos de atenção
- Self-attention (Autoatenção)
- Projeção linear
- Patches de imagen
- Pré-treinamento

Transformers e mecanismos de atenção

- Os Transformers são compostos por blocos de codificador e decodificador que operam em sequências de tokens. Cada bloco contém camadas de autoatenção, que permitem que o modelo capture as relações entre todos os tokens na sequência de entrada.

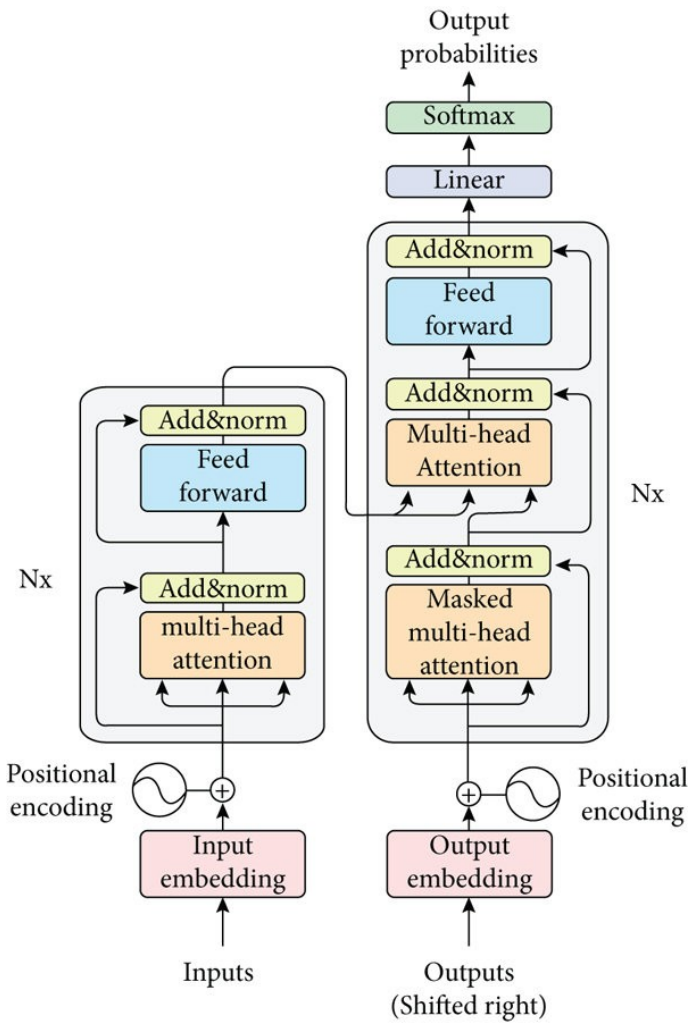
- O mecanismo de atenção calcula pesos para cada par de tokens, permitindo que o modelo se concentre em partes específicas da entrada durante o processamento.

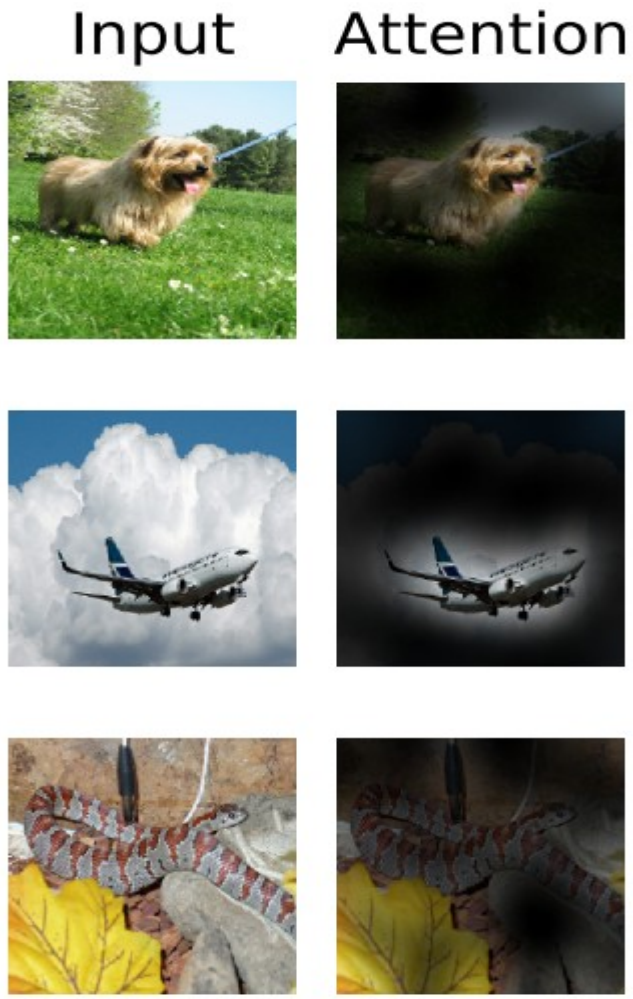
Self-attention

- A autoatenção é um mecanismo que permite que um token em uma sequência "atenda" a outros tokens na mesma sequência, calculando pesos de atenção que indicam a importância relativa de cada token para o token de consulta.

- Isso permite que o modelo capture dependências de longo alcance e aprenda representações contextuais ricas.

Self-attention





Patches de imagen

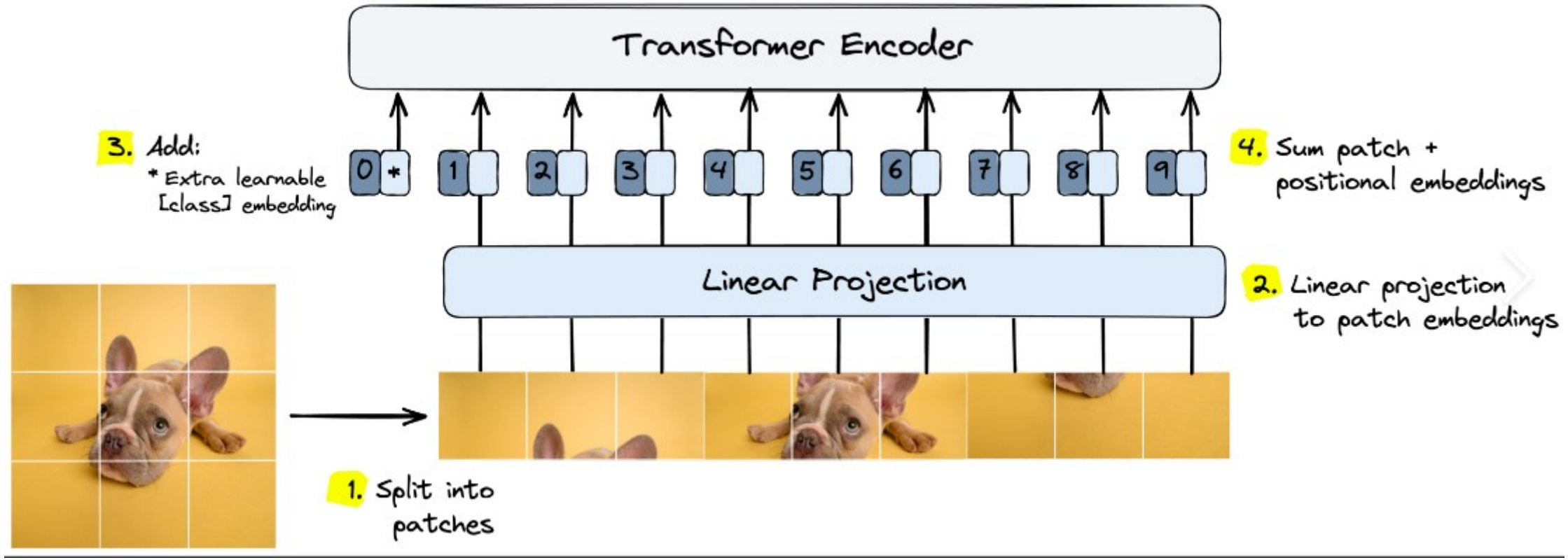
- As imagens de entrada são divididas em patches sobrepostos e cada patch é tratado como um token (palavra) na sequência de entrada do Transformer.

- Isso permite que o modelo processe a imagem como uma sequência de tokens (palavras), capturando informações locais e globais por meio das operações de autoatenção.

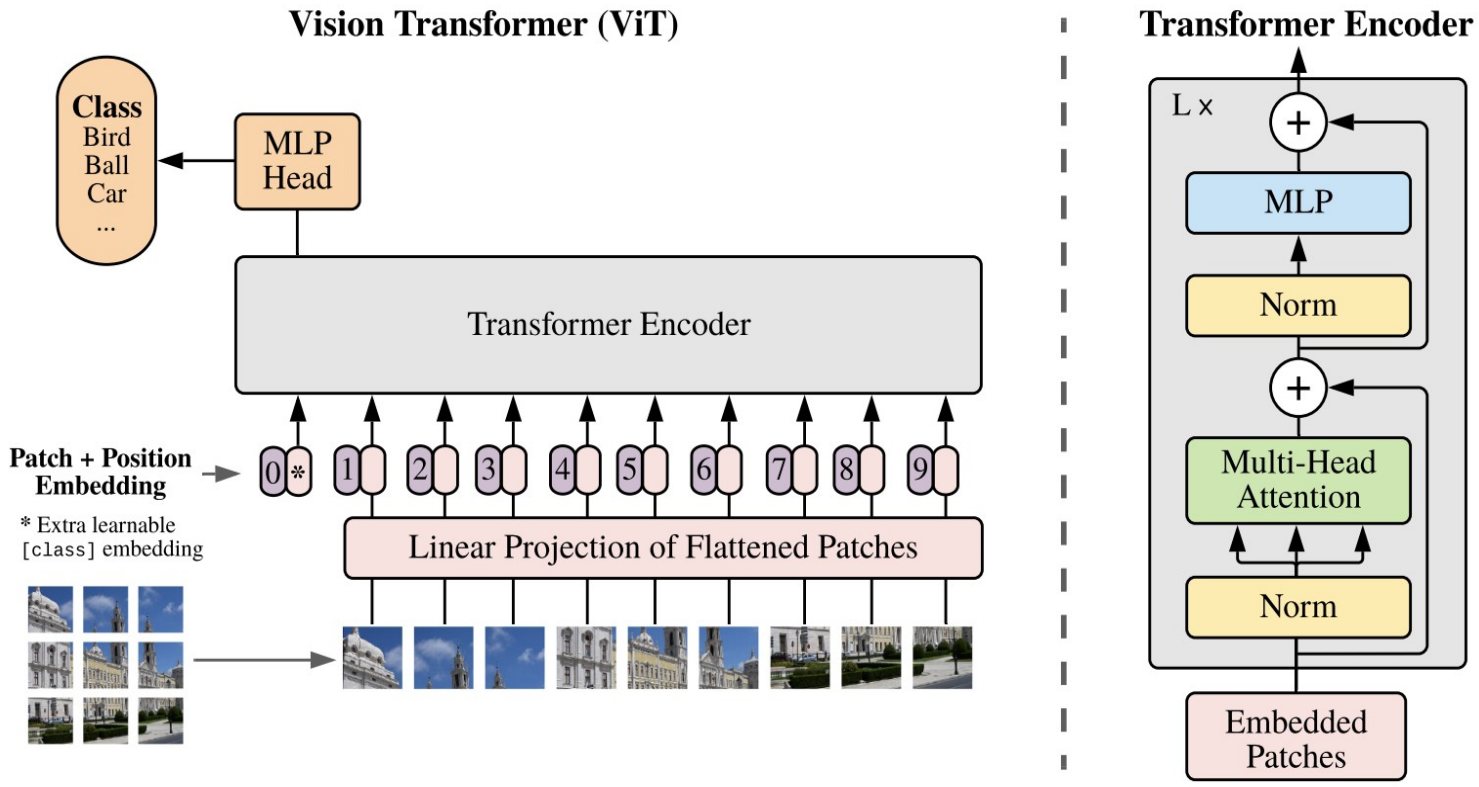
Projeção linear

- A projeção linear é usada para mapear os patches de imagem em um espaço de representação com dimensão constante (D) antes de serem alimentados para o Transformer.
- Essa projeção é realizada por uma camada totalmente conectada treinável, que transforma os patches em vetores de características com a mesma dimensionalidade para garantir consistência ao longo das camadas do Transformer.

Projeção linear



Arquitetura e funcionamento



- Divisão da Imagem em Patches
- Projeção Linear Inicial
- Camadas do Transformer
- Autoatenção e Integração de Informações
- Incorporação de Posição
- Treinamento e Ajuste Fino

Transformer Encorder

- A camada de Transformer Encoder é responsável por processar a sequência de vetores de patches de imagem, aplicando mecanismos de autoatenção e redes neurais totalmente conectadas para capturar informações contextuais e realizar transformações nos dados de entrada.

- Também incorpora mecanismos de normalização e redes neurais totalmente conectadas para processar e transformar os vetores de patches de imagem, permitindo a extração de características relevantes e a representação eficaz das informações visuais ao longo das camadas do modelo ViT.

Treinamento e otimização

O processo de treinamento do Vision Transformer (ViT) envolve as seguintes etapas:

- Pré-Treinamento
- Fine-Tuning
- Aprendizado por Transferência

Treinamento e otimização

Estratégias de Otimização:

- Ajuste de Hiperparâmetros
- Regularização
- Otimizadores
- Ajuste de Taxa de Aprendizado

Vantagens e desvantagens

Vantagens do Vision Transformer (ViT) em relação a outros algoritmos:

- Escalabilidade.
- Eficiência em grandes conjuntos de dados.
- Redução de viéses de arquitetura

Vantagens e desvantagens

Possíveis limitações e desafios associados ao uso do ViT:

- Complexidade computacional.
- Necessidade de grandes conjuntos de dados.
- Interpretabilidade

Exemplo(s) de aplicação

- Reconhecimento de Objetos em Imagens
- Detecção de Anomalias em Imagens
- Detecção de objetos
- Compressão de imagens
- Detecção de vídeo Deepfake
- Análise de cluster
- Classificação de Imagens
- Segmentação de Imagens em Aplicações de Visão Computacional
- Reconhecimento de Padrões em Imagens, etc.

Comparação com outros algoritmos

- **Desempenho Superior em Diversas Tarefas:**
  O ViT atinge excelentes resultados em tarefas de classificação de imagem, superando redes neurais convolucionais (CNNs) em benchmarks como ImageNet, CIFAR-100 e VTAB, enquanto requer menos recursos computacionais para treinar.

Comparação com outros algoritmos

O Vision Transformer (ViT) obtém excelentes resultados quando pré-treinado em escala suficiente e transferido para tarefas com menos pontos de dados. Quando pré-treinado no conjunto de dados público ImageNet-21k ou no conjunto de dados interno JFT-300M, o ViT se aproxima ou supera o estado da arte em vários benchmarks de reconhecimento de imagens. Em particular, o melhor modelo atinge a precisão de 88,55% no ImageNet, 90,72% no ImageNet-ReaL, 94,55% no CIFAR-100 e 77,63% no conjunto VTAB de 19 tarefas.

Comparação com outros algoritmos

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We re-

Perguntas?
Quiz do tema 4
https://forms.gle/wh9aLKZunM3XDmaT7

Referências

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. **An image is worth 16x16 words: Transformers for image recognition at scale, 2021**.

- Abdelhafid Berroukham, Khalid Housni, and Mohammed Lahraichi. Vision transformers: **A review of architecture, applications, and future directions.**In 2023 7th IEEE Congress on Information Science and Technology (CiSt), pages 205–210, 2023. doi:10.1109/CiSt56084.2023.10410015

- Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. **Attention is all you need in speech separation**. In ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 21–25, 2021 doi:10.1109/ICASSP39728.2021.9413901

- Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. **What do vision transformers learn? a visual exploration, 2022**.

- E. Ibrahimovic. **Optimizing vision transformer performance with customizable parameters.** In 2023 46th MIPRO ICT and Electronics Convention (MIPRO), pages 1721–1726, 2023. doi:10.23919/MIPRO57284.2023.10159761.

- Md Sohag Mia, Abu Bakor Hayat Arnob, Abdu Naim, Abdullah Al Bary Voban, and Md Shariful Islam. **Vits are everywhere: A comprehensive study showcasing vision transformers in different domain. In 2023** International Conference on the Cognitive Computing and Complex Data (ICCD), pages 101–117, 2023. doi:10.1109/ICCD59681.2023.10420683.

Obrigado!