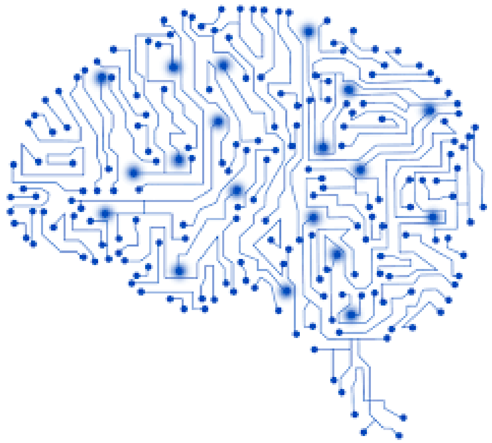
University of Minho
School of Engineering



Aprendizagem Profunda

Patch Embeddings, Vision Transformers, and Vision-Language Models

APP @ MEI/1º ano – 2º Semestre
Victor Alves

Part VIII

Contents

Introduction

Motivation:

•Deep learning in vision traditionally dominated by CNNs.
•Transformers reshaping vision by treating images as sequences.
•Key idea: unifying vision and language architectures via Transformers.

Patch Embeddings

Patch embeddings:

- Divide an image into fixed-size patches (e.g., 16×16 pixels).
- Flatten and linearly project each patch to a vector.

Obs.:

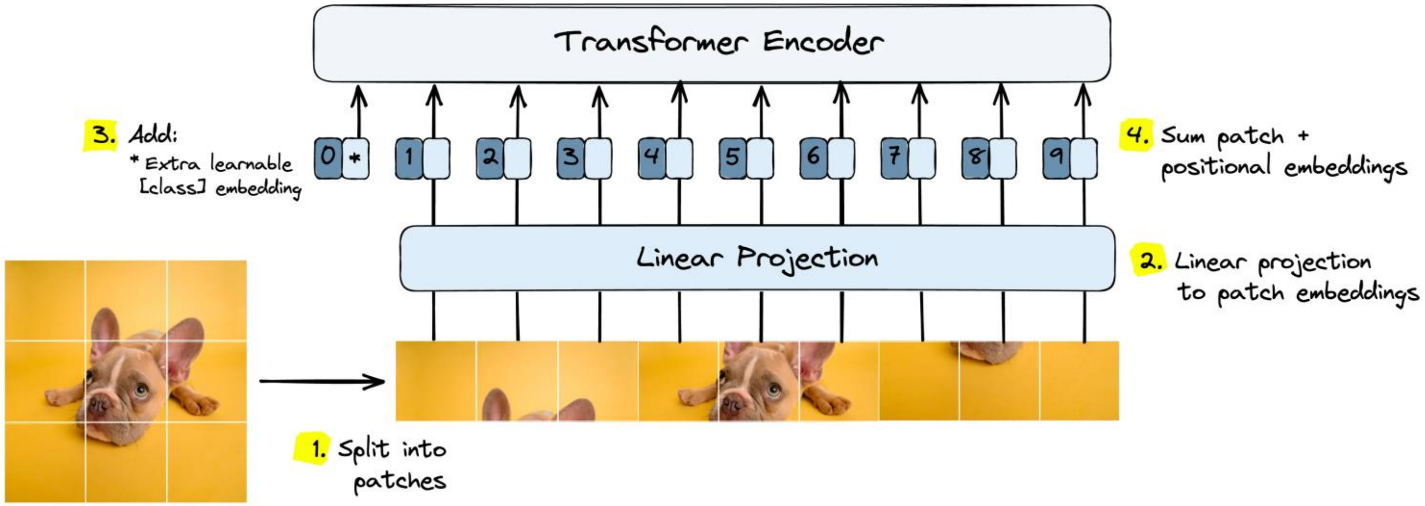
- Transforms 2D image into a 1D sequence → suitable input for a Transformer.

Understanding Image Patch Embeddings

From simple unfolding to 2D convolutions

Nikolaus Correll Follow 9 min read · Feb 4, 2025

https://medium.com/correll-lab/understanding-image-patch-embeddings-3d66c14fe7ed



https://www.pinecone.io/learn/series/image-search/vision-transformers/

Vision Transformers (ViT)

Introduced by Dosovitskiy (A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021), ICLR)

Architecture:

- Patch Embedding + Positional Encoding + Standard Transformer Encoder.
- Classification token ([CLS]) captures global image representation.

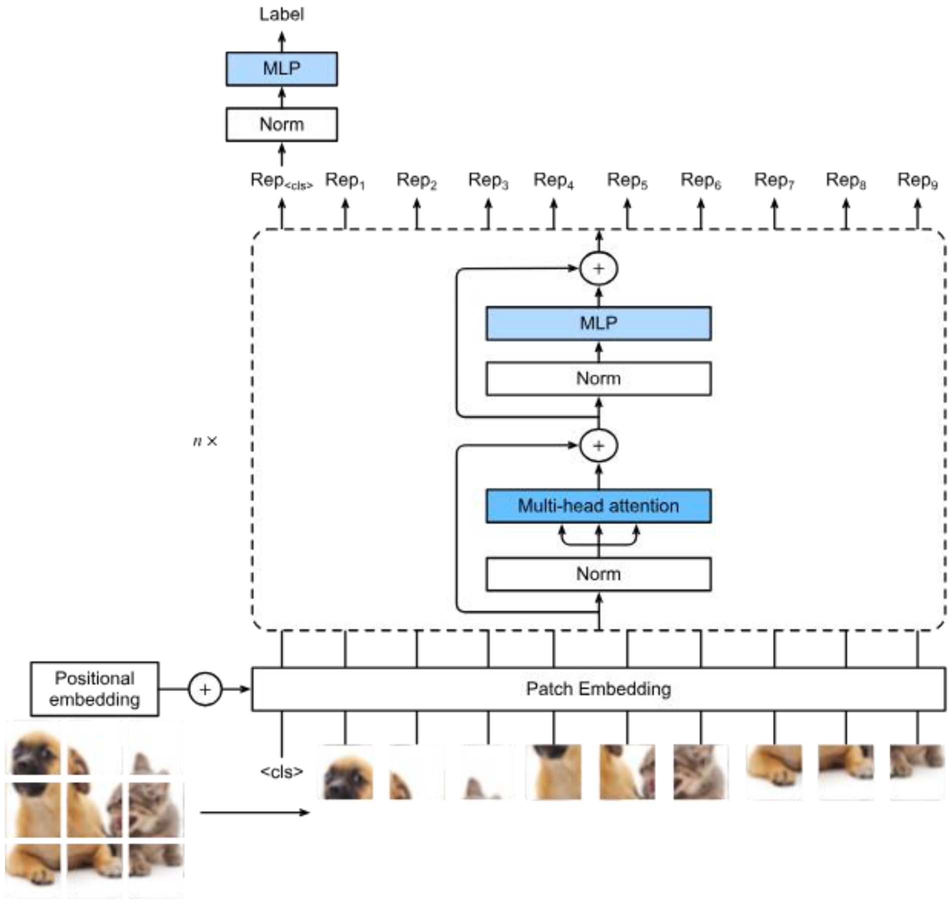
Key Points:

- Scales well with data and compute.
- Competes with or surpasses CNNs when trained on large datasets.
- Can model long-range dependencies and global context in images.
- Flexible with image resolutions and effective in transfer learning scenarios
- Easier to combine with language models.

Use self-attention to capture global relationships between patches, not just local features as in CNNs

CNNs use convolutions to extract local features; ViTs use self-attention for global context.

ViTs often require larger datasets to train effectively but excel in tasks where global relationships are important



Simple ViT block showing patch embedding → transformer layers → classification head. Classification Head: Uses the output (often the class token) for downstream tasks (e.g., image classification)

https://d2l.ai/chapter_attention-mechanisms-and-transformers/vision-transformer.html

Vision-Language Models (VLMs)

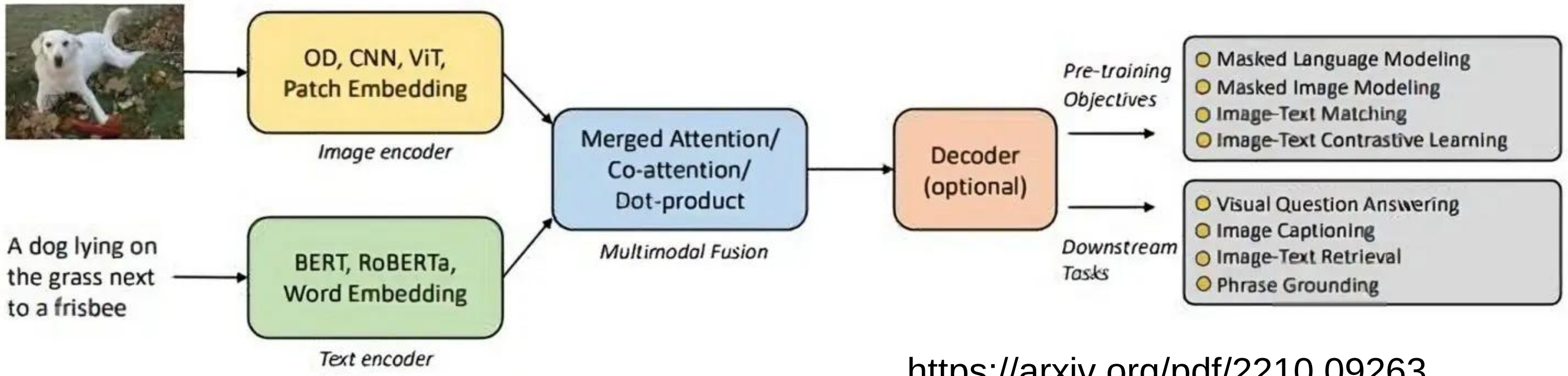
Vision-Language Models:
•VLMs combine computer vision and natural language processing, enabling models to understand and generate both images and text.
•They process images and their textual descriptions together, learning associations between visual and linguistic information.

Architecture components:
- Image Encoder: Extracts visual features from images (often a ViT or CNN).
- Text Encoder: Processes text (often a transformer-based language model)
- Fusion Mechanism: Combines image and text representations for cross-modal understanding.

How they work:
- Both images and text are transformed into embeddings.
- The model learns to align or fuse these embeddings, enabling tasks like generating text from images or answering questions about images

Applications: Image captioning, visual question answering, image-text retrieval, generative AI, …

- DeepSeek-VL2
- Gemini 2.0 Flash
- GPT-4o
- Llama 3.2
- NVLM
- Qwen 2.5-VL



https://arxiv.org/pdf/2210.09263

Challenges

- Alignment between modalities
- Data quality and bias
- Efficient fine-tuning
- Interpretability