Vision Transformer (ViT)

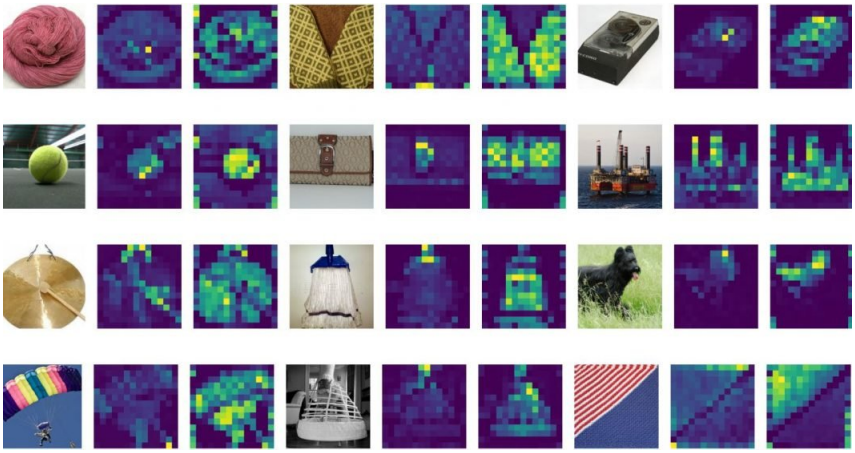
Kramer Roxana, Miruna Sapca, Marian Ostate, Victor Gherghel

West University of Timisoara, Faculty of Mathematics and Computer Science

Abstract.

1 Introduction

Vision Transformer (ViT) is a groundbreaking deep learning architecture that has revolutionized computer vision tasks, departing from traditional convolutional neural networks (CNNs). Introduced by researchers at Google in 2020, ViT leverages the power of transformers, originally designed for natural language processing, to process image data in a highly efficient and scalable manner.



The structure of the vision transformer architecture consists of the following steps:

1. Split an image into patches (fixed sizes)
2. Flatten the image patches
3. Create lower-dimensional linear embeddings from these flattened image patches
4. Include positional embeddings
5. Feed the sequence as an input to a state-of-the-art transformer encoder
6. Pre-train the ViT model with image labels, which is then fully supervised on a big dataset
7. Fine-tune the downstream dataset for image classification

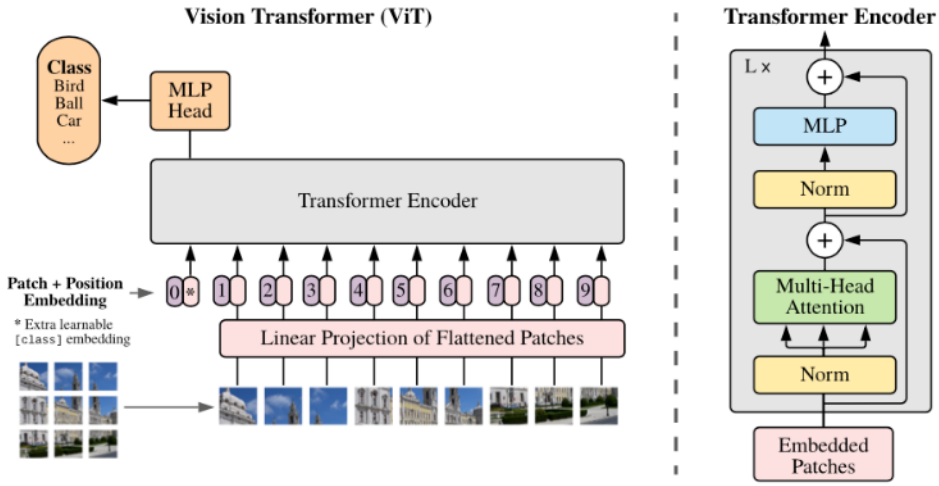


Fig. 1. Vision Transformer ViT Architecture