Vision Transformer (ViT)

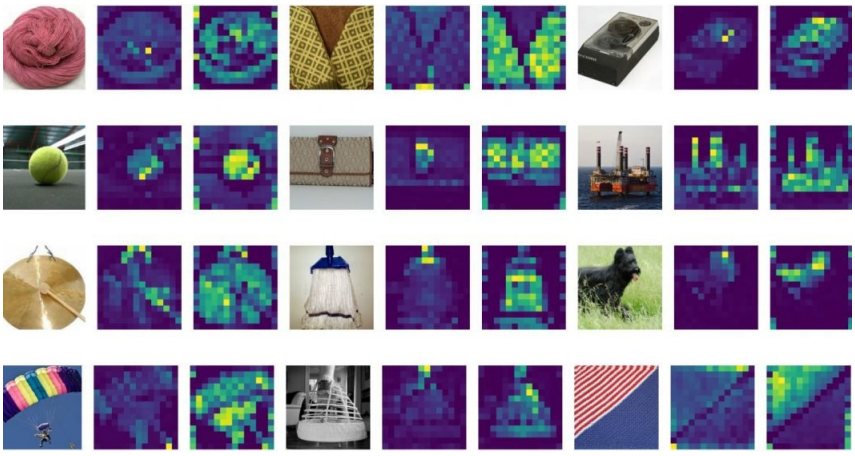
Roxana Kramer, Miruna Sapca, Marian Ostate, Victor Gherghel

West University of Timisoara, Faculty of Mathematics and Computer Science

Abstract. We choosed the Vision Transformer (ViT) model for our project, drawn to its innovative approach and promising capabilities. Our decision to choose ViT comes from a convergence of factors that align perfectly with the distinctive requirements and targets of our project. The big interest and ongoing support within the research community for ViT helped us with making our project decision. This ensures that we are working with a model backed by many collective knowledge. In summary, our choice of ViT is influnced by a colection of its unique features and adaptability to project requirements.

1 Benchmark

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

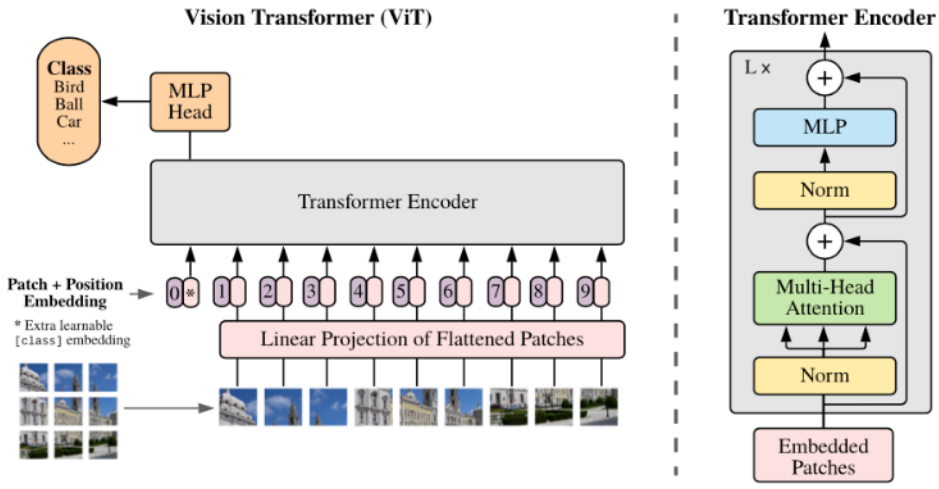


Fig. 1. Vision Transformer ViT Architecture

2 Tools

2.1 alpha-beta-CROWN

$\alpha, \beta-$CROWN functions as a neural network verifier employing a streamlined linear bound propagation framework and branch-and-bound techniques. Its computational efficiency is enhanced when deployed on GPUs, allowing effective

scaling to sizable convolutional networks with millions of parameters. The alpha-beta-CROWN method is capable of offering provable assurances of robustness against adversarial attacks while also verifying other general properties inherent in neural networks.

2.2 PyRat

References

1. Benchmark: https://github.com/ChristopherBrix/vnncomp2023_benchmarks/tree/main/benchmarks/vit
2. Tool 1: https://github.com/Verified-Intelligence/alpha-beta-CROWN
3. Tool 2: https://github.com/ChristopherBrix/vnncomp2023_results/tree/main/pyrat
4. https://viso.ai/deep-learning/vision-transformer-vit/?fbclid=IwAR1lTOAfng_T7diBYAgxWzjSKSpHhNLDBpXwawuCIfRWoP5IMZD1Ufd3GCc