Project 10 Algorithm assignment

m5261113 Akihiro Hattori

m5261176 Yuta Uchida

m5261159 Taisei Shibakura

m5261162 Taisei Shiraishi

**Abstract**

Vision Transformer is an architecture for image recognition proposed in the paper "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale".

When it comes to deep learning in image processing, architectures using convolution based on CNNs have achieved high performance. However, Vision Transformer attracted attention because of its high performance without convolution.

Vision Transformer performed better than the SOTA model without convolution. In addition, ViT-H/14 showed that compared to CNN-based Big Transfer, Vision Transformer is significantly less computationally intensive, by about a factor of 4. This has led to numerous studies of models based on the Vision Transformer today.グラフ, 折れ線グラフ

自動的に生成された説明

Reference: https://paperswithcode.com/method/vision-transformer

**Feature of ViT**

・No convolution

・Update SOTA

・Significantly reduced computation

**Architecture**

Process flow

テキスト

自動的に生成された説明

テーブル が含まれている画像

自動的に生成された説明

Reference: <https://openreview.net/forum?id=YicbFdNTTy>

1, Converts input image to vector

First, the input image is divided into patches as shown in the figure. Then, each of the divided images is converted to a vector.

2, Processed by Transformer Encoder

The vector-transformed image for each batch is then input to the transformer encoder for processing. The processing is shown in the figure on the right.

3, Processed by MLP Head

What you get from the Transformer Encoder can be put into the MLP Head to get the final prediction.