FINAL-REPORT: VISION TRANSFORMER

SongSeungWu

1 WHAT IS VISION TRANSFORMER(VIT)?

Vision Transformer is a model that applies the Transformer architecture to image recognition. Originally successful in natural language processing, the Self-Attention mechanism is extended to computer vision in ViT. Unlike convolutional neural networks, ViT processes images as sequences of patches, similar to how text is tokenized for NLP tasks.

2 KEY FEATURES OF VIT

- Patch Embedding: An image is divided into fixed-size patches, and each patch is converted into an embedding vector to be used as input for the Transformer.
  Images are treated like sequences of tokens.

- Self-Attention Mechanism: ViT models the global relationships between patches using Self-Attention, capturing dependencies across the entire image.
  Unlike CNNs, which focus on local features, ViT can learn global features efficiently in one step.

- Pre-training: ViT is pre-trained on large-scale datasets, achieving high generalization performance.
  Fine-tuning the pre-trained weights makes ViT adaptable to various downstream tasks.

- Tokenization and Positional Embedding: Positional embeddings are added to each patch to encode spatial information since Transformers lack inherent positional awareness.
  Each patch serves as a token for the model.

- Class Token: A special CLS token is prepended to the sequence to summarize the image's overall information.
  The output of this token is used for classification tasks.

3 ADVANTAGES OF VIT

- Global Feature Learning: Self-Attention enables the model to capture relationships across the entire image.

- Efficient Pre-training: ViT leverages large-scale datasets to achieve state-of-the-art performance.

- Modular Architecture: Its Transformer-based structure allows seamless integration with existing NLP models.

4 LIMITATIONS OF VIT

- Data Dependency: ViT requires large datasets to outperform CNNs; on small datasets, it may underperform.

- Computational Cost: The computational complexity of Self-Attention increases rapidly with image resolution and size.

5 VISION TRANSFORMER INFERENCE PIPELINE

The Vision Transformer (ViT) processes input images by dividing them into patches and feeding these patches into a Transformer model for tasks like image classification. The ViT inference pipeline consists of the following steps:

- Dividing the Image into Patches: The 224x224 input image is divided into patches of size 16x16.
  A 2D Convolutional filter is used to produce 14x14 patches.

- Adding Position Embedding: Position information for each patch is encoded using learnable Position Embedding vectors, which are added to the patch embedding vectors.
  This step allows the Transformer to understand the spatial order of the image.

- Transformer Encoder: The patch embedding vectors are fed into the Transformer Encoder.
  The Encoder uses the Self-Attention mechanism to model global relationships between patches, with the input and output vectors having the same dimensions.

- MLP Head: The output of the Transformer Encoder's first CLS token vector is passed to the MLP Head to produce the final classification result.

6 DIVIDING THE IMAGE INTO PATCHES

The Vision Transformer divides the input image into fixed-size patches and converts each patch into a high-dimensional vector to be used as input for the Transformer model.
1. Patch Embedding Process

- The input image is divided into patches of size 16x16, resulting in a total of 14x14 patches.

- Conv2d is used to convert each patch into a 768-dimensional vector.

- The output is structured for input to the Transformer model.

2. PatchEmbed Class Implementation

- The PatchEmbed class uses a Conv2D layer to divide the image into patches and embed each patch into a vector.

7 ADDING POSITION EMBEDDING

Vision Transformer encodes the positional information of each patch using learnable Position Embedding. This helps the model understand the spatial relationships between patches.
1. Adding Position Embedding

- Position Embedding vectors representing spatial information are added to the embedding vectors of each patch.

- The learnable Position Embedding vectors capture the similarity and distance relationships between patches within an image.

2. Visualizing Position Embedding Similarity

- The cosine similarity between one patch and all other patches is computed to visualize the spatial relationships learned by Position Embedding.

3. Generating Transformer Input

- A classification token is prepended to the patch embedding vectors, and Position Embedding is added to form the Transformer input.

8 TRANSFORMER ENCODER

The Vision Transformer employs Transformer Encoders to learn global relationships between image patches. Each Transformer Encoder processes the input vectors through Multi-Head Attention and Fully Connected Layers, maintaining the same dimensionality for inputs and outputs.
Structure of the Transformer Encoder:

- Input: N patch embedding vectors are fed into the Transformer Encoder.
  Each vector contains 768 dimensions, including one token.
- Query, Key, Value Separation: The input vectors are passed through a Fully Connected Layer to produce Query, Key, and Value.
- Multi-Head Attention: Q, K, and V are split into H heads for parallel attention computation.
  The attention outputs are concatenated and reshaped to match the input dimensions.
- Residual Connection & Layer Normalization: A residual connection and layer normalization are applied to stabilize learning.
- MLP: The attention results are passed through two Fully Connected Layers to generate the final Encoder output vectors.

Operation of Multi-Head Attention:

- The input vector is transformed into Query, Key, and Value, and then split into 12 attention heads
- The Attention Matrix is computed using Query and Key
- Each attention head's output can be visualized to interpret the relationships between patches.

9 MLP(CLASSIFICATION) HEAD

The Classification Head in Vision Transformer uses the first output vector, corresponding to the CLS token, from the Transformer Encoder to produce the final classification result.

- Using the CLS Token: The first vector from the Transformer Encoder output represents a summary of the entire image.
  This vector is used as the input to the Classification Head.
- Classification Head: The Classification Head is implemented as a Multi-Layer Perceptron.
- Classification Result: The final output vector is processed with argmax to select the class ID with the highest probability.

10 CONCLUSION

The Vision Transformer successfully processed the input image through its pipeline, from patch embedding to classification, leveraging the Transformer architecture. The experimental results demonstrate that ViT accurately predicted the input image as church, church_building.
ViT showcases the successful extension of Transformer architecture to image classification. Unlike traditional CNN-based approaches, ViT excels at learning relationships between patches, resulting in outstanding performance.