Nama: Rizki Hidayat

NIM: 1103202131

Laporan

1.Pendahuluan

Pada modul ini, mengimplementasikan arsitektur Vision Transformer (ViT) sebagai model pengenalan gambar. Tujuannya adalah untuk memvalidasi temuan dari paper asli, memastikan reproduktibilitas, dan mendapatkan pemahaman yang lebih dalam tentang perilaku model tersebut.

2.Penjelasan

2.1. Setup Library dan Versi PyTorch

- Menyertakan library yang diperlukan dan memastikan versi PyTorch dan torchvision yang sesuai.

2.2 Pemrosesan Data

- Mengunduh dan mempersiapkan dataset gambar berisi pizza, steak, dan sushi.
- Menetapkan ukuran gambar dan membuat pipeline transformasi.

2.3 Pembuatan Patches dari Gambar

- Menggunakan modul Conv2d untuk mengubah gambar menjadi sekuens patch.
- Menunjukkan visualisasi patch pada gambar.

2.4 Pembuatan Blok Embedding

- Membuat kelas PatchEmbedding yang merinci langkah-langkah embedding untuk patch.
- Melakukan percobaan menggunakan gambar dan memeriksa bentuk output.

2.5. Penambahan Token Kelas dan Positional Embeddings

- Menambahkan token kelas dan positional embeddings ke patch embeddings.

2.6 Blok Multihead Self-Attention (MSA)

- Membuat kelas MultiheadSelfAttentionBlock yang menerapkan blok MSA dari ViT.

2.7 Blok Multi-Layer Perceptron (MLP)

- Membuat kelas MLPBlock yang menerapkan blok MLP dari ViT.

2.8 Blok Encoder Transformer

- Membuat kelas TransformerEncoderBlock yang menggabungkan MSA dan MLP blocks.

2.9 Model Utama Vision Transformer (ViT)

- Membuat kelas ViT yang menyusun seluruh arsitektur ViT.
- Menyertakan blok MSA, MLP, dan encoder Transformer.
- Menambahkan token kelas, positional embeddings, dan classifier head.

2.10 Pembuatan dan Pemrosesan Data

- Mendemonstrasikan pembuatan token kelas, ekspansi ke dimensi batch, dan penggunaan ViT pada gambar-gambar acak.
- Menggunakan class ViT yang diimplementasikan sebelumnya untuk pembuatan model ViT.

2.11 Pelatihan Model ViT dari Awal

- Membuat objek optimizer dan fungsi kerugian untuk pelatihan model.
- Menggunakan modul engine untuk melatih model ViT dari awal selama 10 epoch.
- Menampilkan grafik kurva kerugian hasil pelatihan.

2.12 Penggunaan Model ViT Pre-trained

- Menggunakan pre-trained weights dari ViT-Base menggunakan torchvision.
- Mengganti lapisan classifier head dan membekukan parameter base model.
- Melatih classifier head yang baru pada dataset gambar pizza, steak, dan sushi.
- Menampilkan grafik kurva kerugian hasil pelatihan classifier head baru.
- Menyimpan model hasil pelatihan.

2.13 Evaluasi Model

- Membuat prediksi pada gambar pizza menggunakan model yang telah dilatih dan menampilkan hasilnya.