Google Developer Experts

Diabetic Retinopathy Grade Classification using Vision Transformers

Workshop

Sara EL-ATEIF
ML GDE
@el_ateifSara

About Me

Google Ph.D. Fellow

Pursuing my Ph.D. @ENSIAS, Morocco

NVIDIA DLI Instructor/University Ambassador

Co-Founder of AI Wonder Girls

Experts

Overview

Proprietary + Confidential

Plan

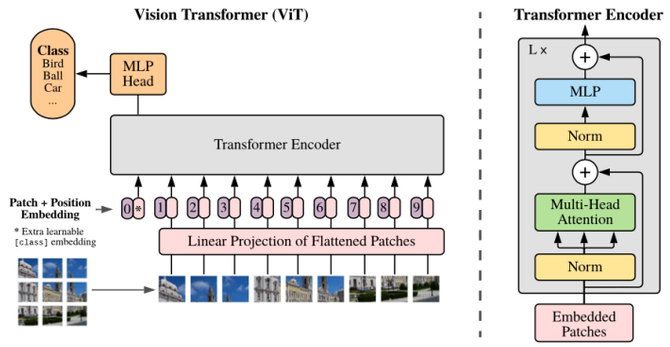
- Vision Transformers
- CNN vs Vision Transformers
- Diabetic Retinopathy
- Workshop

Vision Transformers

About ViT

The Vision Transformer, or ViT, image classification model that employs a Transformer-like architecture over patches of the image.
(1) An image is split into fixed-size patches,
(2) Each image is then linearly embedded,
(3) Position embeddings are added,
And the resulting sequence of vectors is fed to a standard Transformer encoder.

To perform classification, an extra learnable “classification token” is added to the sequence.



Source: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (paper by Dosovitskiy et al.)

Source: Lilian Weng

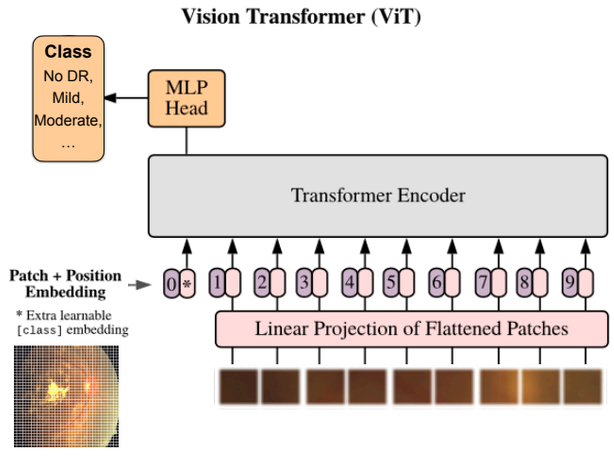
Source: Attention Is All You Need

Experts

- The Vision Transformer, or ViT, is a model for image classification that employs a Transformer-like architecture over patches of the image. An image is split into fixed-size patches, each of them are then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder. In order to perform classification, the standard approach of adding an extra learnable “classification token” to the sequence is used.
- Source: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (paper by Dosovitskiy et al.)

Introduction

Architecture



Source: Lilian Weng

Source: Attention Is All You Need

Experts

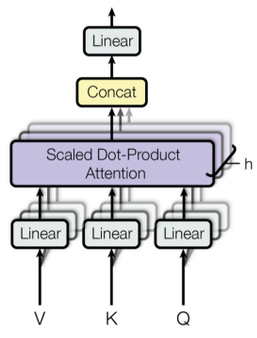
- Talk about How ViT splits images into patches and adds positional embeddings etc

Introduction

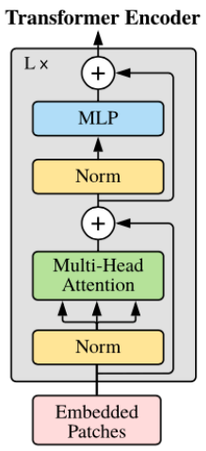
Transformer Encoder part

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \ldots, \text{head}_h]\mathbf{W}_0$$

$$\text{where head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right)$$



Multi-Head Attention



Source: Lilian Weng

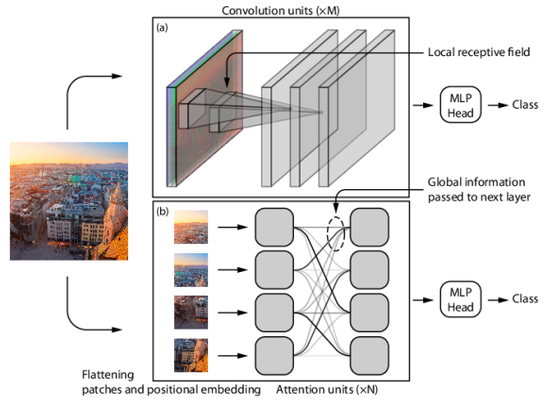
Source: Attention Is All You Need

Experts

- Multi-head Attention is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. Intuitively, multiple attention heads allows for attending to parts of the sequence differently (e.g. longer-term dependencies versus shorter-term dependencies).
- Source: Lilian Weng
- Source: Attention Is All You Need

CNN vs ViT

Image representation



Source: Are Convolutional Neural Networks or Transformers more like human vision?
Shikhar Tuli et al.

CNNs vs ViTs

CNNs:
- Focus on local information
- Needs more training time
- Lots of resources
- Vulnerable to adversarial attacks or changes in the data

ViTs:
- Focus on global information
- Needs less training time
- Resources friendly
- Robust against adversarial attacks

Diabetic Retinopathy

Definition

Diabetic Retinopathy worldwide

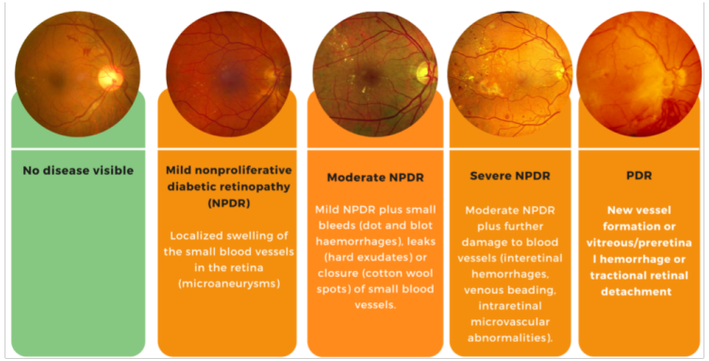
‘Diabetic Retinopathy (DR) is a complication of diabetes, caused by high blood sugar levels damaging the back of the eye (retina). It can cause blindness if left undiagnosed and untreated.’

Source: NHS UK

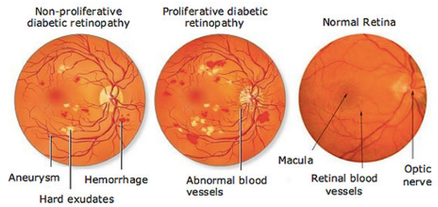
‘Globally, the number of people with DR will grow from 126.6 million in 2010 to 191.0 million by 2030.’

Source: 10.4103/0301-4738.100542

Grades & Symptoms



Source: https://www.ophthalytics.com/our-technology/diabetic-retinopathy/

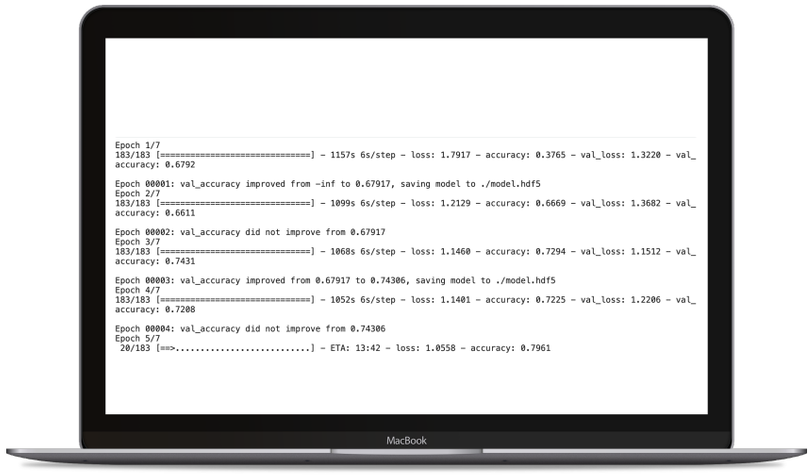


Source: https://lowvisionaids.org/diabetic-retinopathy/

Experts

Proprietary + Confidential

Let’s see the code



Experts

Thank You!

Sara EL-ATEIF
ML GDE
@el_ateifSara