1. Vision Transformer

1.1 Revision of Vanilla Transformer

Mô hình Vision Transformer (ViT) là một biến thể của kiến trúc Transformer được tối ưu hóa cho xử lý hình ảnh. Dưới đây là những điểm khác biệt chính giữa ViT và Vanilla Transformer:

1. Kiểu dữ liệu đầu vào:

* Transformer gốc (NLP):

Đầu vào: Dữ liệu chuỗi văn bản, thường là danh sách các từ hoặc các mã hóa của chúng (ví dụ: [1, 5, 27, 9]).

* ViT (Vision Transformer):

Đầu vào: Hình ảnh, thường là tensor hình ảnh 2D được chia thành các patch và biểu diễn dưới dạng tensor (ví dụ: kích thước [C, H, W] cho hình ảnh màu).

1. Kiến trúc chú ý (Attention Mechanism):

* Transformer gốc (NLP):

Sử dụng tự chú ý (self-attention) để xử lý thông tin trong các chuỗi văn bản. Mỗi từ có khả năng tương tác với tất cả các từ khác.

* ViT (Vision Transformer):

Sử dụng tự chú ý đa đầu (multi-head self-attention) để xử lý thông tin trong các patch hình ảnh. Các patch có khả năng tương tác với nhau trong không gian 2D.

1. Các lớp mạng nơ-ron:

* Transformer gốc (NLP):

Sử dụng mạng nơ-ron truyền thẳng (feed-forward neural network) sau mỗi lớp tự chú ý để biến đổi thông tin trong từng vị trí của chuỗi văn bản.

* ViT (Vision Transformer):

Sử dụng mạng nơ-ron truyền thẳng (feed-forward neural network) sau mỗi lớp tự chú ý để biến đổi thông tin trong từng patch hình ảnh.

1. Tích hợp thông tin không gian:

* Transformer gốc (NLP):

Xử lý thông tin theo chiều thời gian trong chuỗi văn bản.

* ViT (Vision Transformer):

Xử lý thông tin không gian trong hình ảnh, cho phép hiểu cấu trúc không gian của đối tượng trong hình ảnh.

1. Phân loại đối tượng:

* Transformer gốc (NLP):

Thường được sử dụng để dự đoán và sinh chuỗi văn bản (dịch máy, tạo văn bản).

* ViT (Vision Transformer):

Thường được sử dụng để phân loại đối tượng trong hình ảnh thành các lớp cụ thể.

1. Ứng dụng:

* Transformer gốc (NLP):

Sử dụng cho xử lý ngôn ngữ tự nhiên, dịch máy, tạo văn bản tự động và nhiều tác vụ khác liên quan đến ngôn ngữ.

* ViT (Vision Transformer):

Sử dụng cho xử lý hình ảnh, bao gồm phân loại hình ảnh, phát hiện đối tượng, và nhiều ứng dụng xử lý hình ảnh khác.

Tóm lại, Vision Transformer (ViT) và mô hình Transformer gốc cùng một kiến trúc gốc Transformer, nhưng đã được tối ưu hóa và thích nghi cho các loại dữ liệu và nhiệm vụ khác nhau. ViT tập trung vào xử lý hình ảnh và phân loại, trong khi mô hình Transformer gốc thường được sử dụng cho xử lý ngôn ngữ tự nhiên và tác vụ liên quan đến ngôn ngữ.

1.2 Vision Transformers Architecture

Vision Transformer (ViT) là một mô hình thị giác máy tính dựa trên kiến trúc Transformer, được sử dụng cho các nhiệm vụ như phân loại hình ảnh, phát hiện đối tượng, và phân đoạn hình ảnh. Mô hình này xử lý dữ liệu bằng cách chia chúng thành các patch và sử dụng cơ chế tự chú ý để hiểu quan hệ giữa các patch. Cấu trúc chính của mô hình ViT bao gồm:

- Phân tích dữ liệu thành các Patch:

Đầu tiên, dữ liệu đầu vào được chia thành các patch dữ liệu có kích thước cố định. Mỗi patch được coi là một phần của dữ liệu và được biểu diễn dưới dạng vector.

- Chuyển đổi Patch thành Vector Nhúng (Embedding):

Mỗi patch dữ liệu sau đó được biến đổi thành một vector nhúng (embedding vector) thông qua một lớp chuyển đổi (ví dụ: lớp tích chập). Các vector nhúng này chứa thông tin về mỗi patch và sẽ được sử dụng làm đầu vào cho mô hình.

- Sử dụng Self-Attention cho mối quan hệ toàn cục:

Đặc điểm quan trọng của ViT là việc sử dụng cơ chế tự chú ý (self-attention) để xác định mối quan hệ giữa các patch. Cơ chế tự chú ý giúp mô hình biết được patch nào quan trọng và patch nào cần được xem xét khi dự đoán lớp của dữ liệu.

Bằng cách này, mô hình có khả năng nắm bắt các mối quan hệ xa hơn giữa các phần khác nhau của dữ liệu, điều này giúp cải thiện khả năng hiểu và phân loại dữ liệu.

- Mô hình Transformer Encoder:

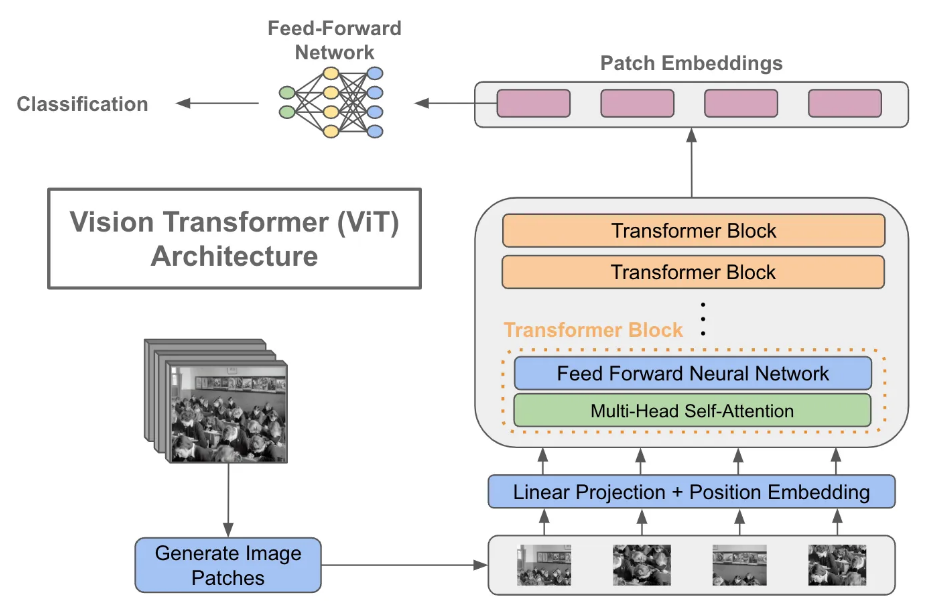
Các vector nhúng của các patch được đưa vào một mô hình Transformer Encoder, cụ thể là một số lớp Transformer Encoder Layers. Mỗi lớp Encoder chứa các lớp tự chú ý và lớp kết nối đầy đủ (fully connected) để xử lý dữ liệu.

Một số lớp Encoder được xếp chồng lên nhau để tạo ra sự biểu diễn phức tạp và có khả năng học các mẫu phức tạp trong dữ liệu.

- Áp dụng lớp Fully Connected Layer:

Sau khi thông tin từ các patch đã được truyền qua một loạt các lớp Encoder, một lớp kết nối đầy đủ cuối cùng được áp dụng để phân loại dữ liệu thành các lớp khác nhau.

Đầu ra của lớp này là các xác suất cho mỗi lớp phân loại.



A basic depiction of a vision transformer architecture

2. Imbalanced Data

2.1 Hàm mất mát Focal Loss

Focal Loss (Hàm Mất Mát Focal) là một hàm mất mát được thiết kế để xử lý vấn đề mất cân bằng lớp (class imbalance) trong các nhiệm vụ phân loại trong học máy và học sâu.

Sử dụng Focal Loss giúp giảm thiểu tác động của mất cân bằng lớp và tập trung vào việc cải thiện khả năng phân loại cho các dữ liệu mất cân bằng. Điều này có thể dẫn đến việc huấn luyện mô hình có độ hiệu suất tốt hơn trên các lớp thiểu số và tỷ lệ false positive (dự đoán sai) thấp hơn.

Hàm mất mát Focal Loss được tính bằng công thức sau đây:

Trong đó:

​ là xác suất dự đoán của mô hình cho lớp đúng.

​ là một hệ số cân bằng để kiểm soát đóng góp của từng lớp.

γ là một tham số để điều chỉnh độ tập trung vào các dữ liệu phức tạp hơn.

2.2 Generative Method