**VGG19 model architecture**

A picture containing text, diagram, plan, technical drawing

Description automatically generated

Đầu tiên, 1 2-D convolutional layer sẽ được áp dụng riêng chi từng ảnh input với hàm activation ReLU để trích xuất các đặc trưng. Layer này có 64 filter, một kernal có kích thước 3x3, theo sau bởi một convolutional layer khác có 64 filter với hàm activation ReLU. Để làm cho đầu ra của convolution ít phức tạp hơn, một MPL với kích thước ma trận là 2x2 được sử dụng.

Kế đến là 2 convolutional layer với 128 filter, kernal 3x3 và activation là ReLU. Các lớp này cho phép VGG19 phân biệt các đặc trưng cấp cao hơn có thể bị bỏ sót trong conv layer trước đó.

Thứ ba là một lớp MPL có kích thước 2x2 và theo sau là 4 lớp 2-D conv-layer với 256 filter, kernel 3x3. Tiếp theo là một MPL và 4 2-D conv layer với 512 filter và thêm một MPL.

Thứ năm, 4 2-D conv layer khác có cấu hình gồm 1024 filter, theo sau là một MPL.

Thứ sáu, hai Fully Connected Layer được cấu hình với 4096 neural và activation ReLU, tiếp theo là một fully connected layer khác với 1000 neural

Cuối cùng, đầu ra được giảm xuống chỉ còn 14 lớp bằng cách áp dụng activation softmax.

**EfficientNetV2-S architecture**

A picture containing text, screenshot, font, number

Description automatically generated

A picture containing text, diagram

Description automatically generated

EfficientNetV2 là một phiên bản cải tiến cảu EfficientNetV1 nhằm mục đích tăng tốc quá trình training và nâng cao hiệu quả tham số. Tốc độ training nhanh nhất đạt được khi sử dụng cả hai convolutional block mới là MBConv và Fused-MBConv.

MBConv là cấu trúc cơ bản của MobileNetV2. Khác biệt về cấu trúc của MbConv và Fused-MbConv là hai block cuối cùng. Trong khi MBConv sử dụng depthwise convolution (3x3) theo sau bởi 1x1 convolution layer, thì Fused-MBConv thay thế 2 layer cuối với 1 3x3 convolutional layer.

Các Fused MBConv layer có thể giúp train nhanh hơn chỉ với một lượng nhỏ tham số tăng lên, nhưng nếu nhiều block trong số này được sử dụng, nó có thể làm chậm quá trình train với nhiều tham số được thêm vào. Để khắc phục vấn đề này, các tác giả đã sử dụng cả MBConv và Fused-MBConv trong neural architecture search, tự động quyết định sự kết hợp tốt nhất của các block này để có hiệu suất và tốc độ tốt nhất.

**ConvNeXt:**

ConvNeXt là một mô hình SOTA đã đạt được kết quả tốt nhất là 87.8% trên bộ ImageNet. Mô hình được xây dựng hoàn toàn từ mạng CNN truyền thống, kết quả thu được tốt hơn Transformer về độ chính xác và khả năng mở rộng trong khi vẫn duy trì tính đơn giản và hiệu quả của mạng CNN.

Về kiến ​​trúc mạng, ConvNeXt phát triển từ mạng ResNet ban đầu và từng bước cải thiện mô hình bằng cách mượn thiết kế của Swin Transformer.

Hình {} thể hiện 4 giai đoạn của ConvNeXt. Tỉ lệ mỗi block trong mỗi giai đoạn là 1:1:9:1.

Mạng CNN truyền thống thường xử dụng Batch Norm và activation là ReLu, trong khi đó ConvNeXt sử dụng Layer Norm và activation là GeLu.

