**Transformers:** Set vectorize data

Transformer models work by processing input data, which can be sequences of tokens or other structured data, through a series of layers that contain self-attention mechanisms and feedforward neural networks

Transformers utilize self-attention mechanisms to capture global dependencies and relations between image patches directly

---

REFERENCE: https://spotintelligence.com/2023/10/13/pre-trained-models/

## Understanding Pre-Trained Models

Pre-trained models have become a game-changer in artificial intelligence and machine learning. They offer a shortcut to developing highly capable models for various tasks, from natural language understanding to computer vision.

Pre-trained models are neural network architectures that have undergone a two-step process: **pre-training and fine-tuning**.

Pre-trained models have garnered immense attention and have become a driving force in many machine learning applications

**Natural Language Processing (NLP)**

**BERT (Bidirectional Encoder Representations from Transformers)** – used for sentiment analysis, text classification, and question answering.

**GPT-3 (Generative Pre-trained Transformer 3)** – can generate human-like text for various tasks, from writing articles to composing poetry.

**XLNet** – leverages a permutation-based training approach

**Computer Vision**

**VGG16 and VGG19:** The **Visual Geometry Group (VGG) models**, with 16 and 19 layers – wisely used for image classification and object recognition.

**ResNet (Residual Network)** – improved training deep neural networks with its deep residual learning framework + renowned for its ability to tackle the vanishing gradient problem. This makes it a go-to choice for image classification and object detection.

**Inception** – known for their innovative architecture featuring inception modules. They are well-suited for image classification and object recognition tasks.

## Audio and Speech Recognition

**Wav2Vec 2.0** – pre-trained model for automatic speech recognition (ASR), crucial for applications like transcription services and voice assistants.

**DeepSpeech** – open-source ASR engine based on deep learning. It's designed for robust and accurate speech recognition, making it an important pre-trained model for speech-related applications.

## Getting Started with Pre-Trained Models

- Choose a framework
- Explore pre-trained model repositories
- Understand the model architecture
- Model selection
- Installation and loading
- Data preparation
- Fine tuning (optional)
- Inference and evaluation
- Deployment
- Ongoing monitoring and updates
- Ethical and regulatory considerations

## PyTorch Pre-Trained Models

***Image Classification*:**

- ResNet (Residual Network) – ResNet-18, ResNet-34, ResNet-50 etc
- VGG models – VGG16 and VGG19
- AlexNet
- DenseNet

***Object Detection*:**

- Faster R-CNN
- YOLO (You Only Look Once)

***Semantic Segmentation*:**

- FCN (Fully Convolutional Network)
- U-Net

***Style Transfer*:**

- VGG-19 with Batch Normalization

***Text Detection***

- EAST (Efficient and Accurate Scene Text Detector)

***Super-Resolution***

- ESRGAN (Enhanced Super-Resolution Generative Adversarial Network)

**TensorFlow Pre-Trained Models**

***Image Classification:***

- Inception – Inception V3 and Inception ResNet.
- MobileNet
- ResNet – ResNet-50, ResNet-101, and ResNet-152.

***Object Detection:***

- SSD (Single Shot MultiBox Detector):
- Faster R-CNN

***Text Generation***:

- GPT-2

***Style Transfer:***

- Arbitrary Image Stylization

***Speech Recognition:***

- Wav2Vec 2.0

***Super-Resolution:***

- ESRGAN (Enhanced Super-Resolution Generative Adversarial Network)

## Imagenet:

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale. One high level motivation is to allow researchers to compare progress in detection across a wider variety of objects -- taking advantage of the quite expensive labeling effort.

## ResNet Model

A Residual Neural Network (ResNet) is an Artificial Neural Network (ANN) of a kind that stacks residual blocks on top of each other to form a network.

There are many variants of ResNet architecture i.e. same concept but with a different number of layers. We have ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-110, ResNet-152, ResNet-164, ResNet-1202 etc.

### Variants of ResNet architecture

Some commonly used variants of resnet architecture are as follows:

- **ResNet-18:** Contains 18 layers and is often used for smaller datasets or tasks where computational resources are limited.
- **ResNet-34:** A deeper version, containing 34 layers.
- **ResNet-50:** One of the most commonly used variants, with 50 layers. It is often a good balance between depth and computational efficiency.
- It is a convolutional neural network (CNN) that excels at image classification. It's like a highly trained image analyst who can dissect a picture, identify objects and scenes within it, and categorize them accordingly.
- **ResNet-101 and ResNet-152:** These models go even deeper and are used in more complex tasks where the highest possible accuracy is needed.

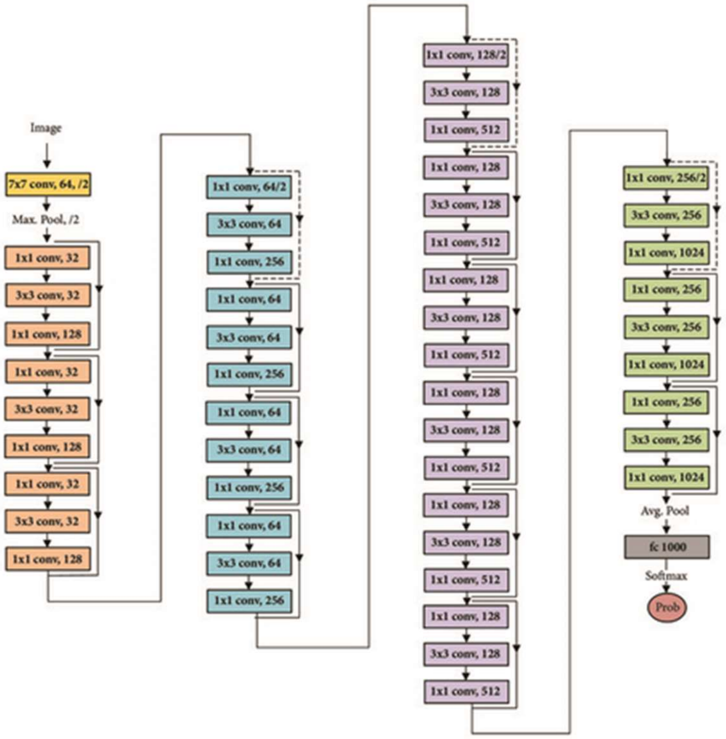

Image: RESNET50 Model

**To find pre-trained models**

REFERENCE: https://keras.io/api/applications/

## VGG Model

VGG (Visual Geometry Group) is a convolutional neural network architecture that was proposed by researchers from the University of Oxford in 2014. It gained popularity and recognition for its simplicity and effectiveness in image classification tasks.

- Convolutional Layers: The architecture uses small 3x3 convolutional filters throughout the network. All convolutional layers in VGG use the same kernel size (3x3) and padding (1 pixel).
- After a series of convolutional and pooling layers, VGG models include three fully connected layers at the end, followed by a softmax layer for classification.
- The first two fully connected layers have 4096 units each, and the final one has 1000 units corresponding to the ImageNet classes.
- VGG uses max-pooling layers with a 2x2 filter and a stride of 2, following some of the convolutional layers. This reduces the spatial dimensions of the feature maps while retaining the important features.

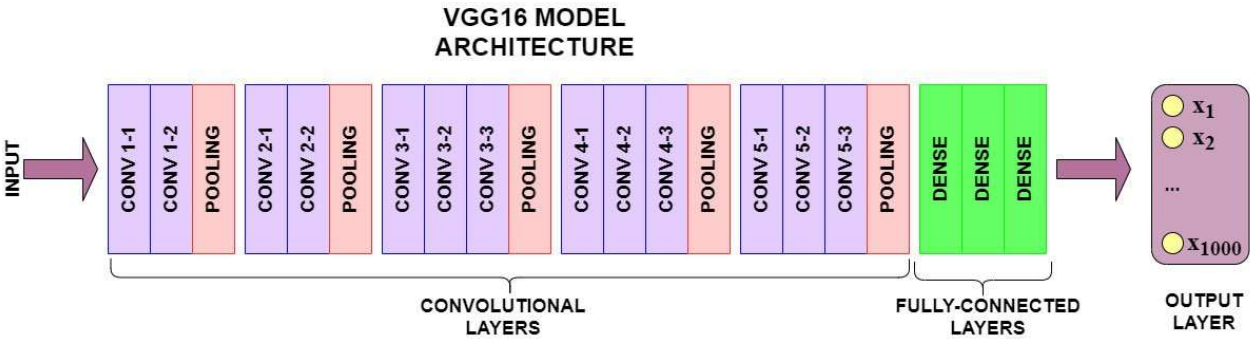

Image: VGG16 Model

**Model Variants:**

- VGG-11: A shallower version with 11 layers.
- VGG-13: Contains 13 layers, with a few more convolutional layers compared to VGG-11.
- VGG-16: The most popular variant with 16 layers (13 convolutional layers and 3 fully connected layers).
- VGG-19: The deepest variant with 19 layers (16 convolutional layers and 3 fully connected layers).