Theoretical aspects of project implementation

AUTOMATED TAGGING AND DESCRIPTION FOR ECOMMERCE PRODUCTS

TEAM MEMBERS:
VASUDHA RANI PATHEDA-(21MIS7121)
GALETI YASWANTH SAI-(21BCE9674)
S CHIRANJEEVI-(21BCE8950)

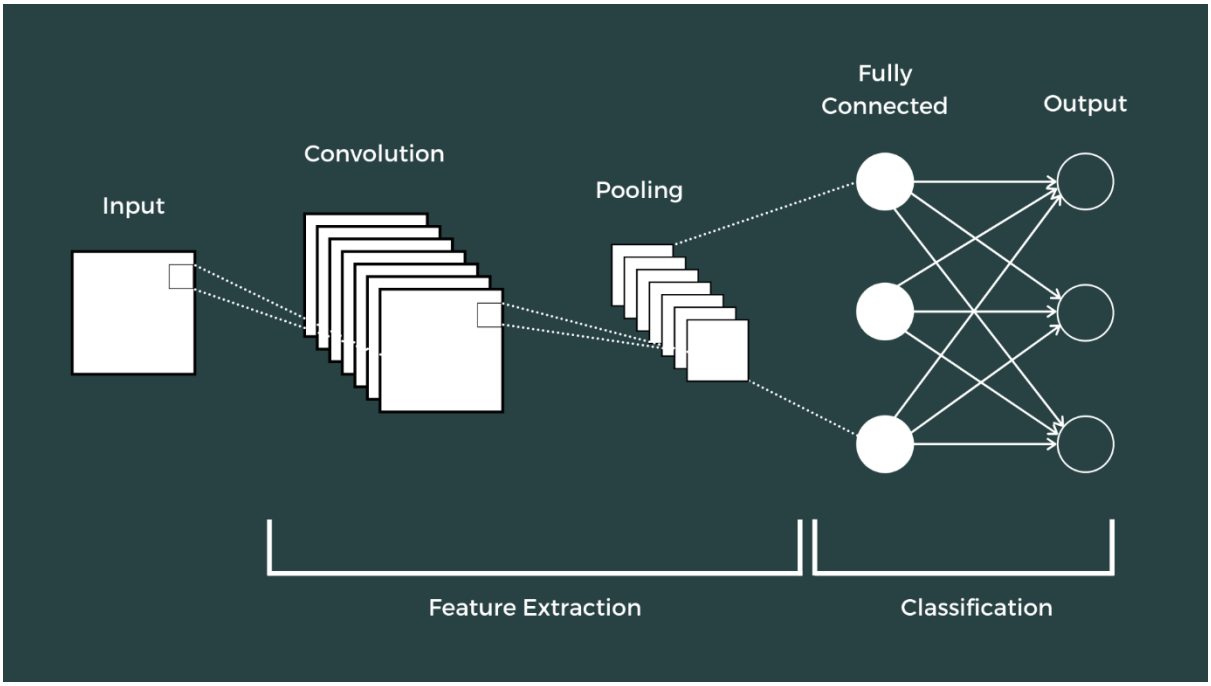
Comparison of three models used in our e-commerce image classification project. Two of these models-CNN and VGG16-are already renowned within the area of image classification, while the third is the newly proposed hybrid model. The objective of this work is to compare and analyse each of the models based on their stated accuracy, efficiency processing power, and their ability to classify product categories ranging from appliances, clothing to electronics in the e-commerce domain.

The e-commerce dataset contains approximately 3 lakh images, with labels of product name, main category, and subcategory. The task is the classification of these images into their respective product categories on the basis of visual attributes like shape, colour, and text on products.

1. **Convolutional Neural Network (CNN)**

   Deep learning models have received much popularity in the area of image classification - the Convolutional Neural Network, or CNN, for example. A CNN is specifically focused on processing and classifying images, recognizing edges, shapes, and textures that make up images. The model works through a combination of images with convolution filters, extracting the spatial hierarchy along with important features of a given image. This model is very effective when the goal is to extract visual features at various levels, so it stands a good chance in image classification applications in e-commerce.

   Mostly, the CNN is applied in classifying products through characteristic images such as shape, color, and text patterns. An example of product classification is appliances, which may be categorized as air conditioners through distinguishing kinds of buttons, logos, or several other kinds of displays.
   Key Components:
   Input Layer- It takes the raw image data as input into the network.
   Convolutional Layers: Extract visual features using convolution filters. Each layer captures complex patterns, edges, and textures, so on.
   Pooling Layers: It reduces the dimensionality of the feature map by discarding redundant information and, therefore, increases efficiency in computation.
   Fully Connected Layers: Compile the features extracted into making predictions with regard to the class of the image.
   Activation Functions (ReLU): Introduce non-linearity into the model allowing it to learn more complex patterns within the data.



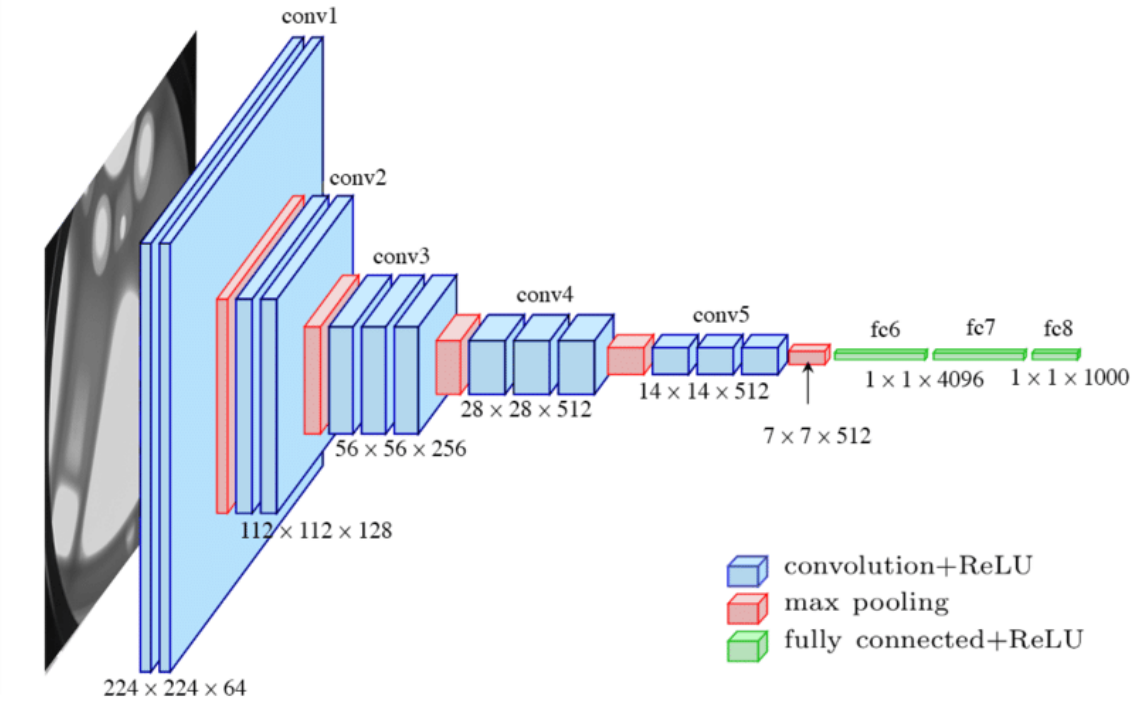
2. VGG16:

VGG16 is perhaps the most widely used deep-learning architectures. Simplicity and capability to handle image classification tasks made it an ideal combination. It contains 16 layers comprising several convolutional layers followed by max-pooling layers. The model uses a smaller size of filters with increased depth by stacking several convolutional layers instead of using larger filters, which give more capabilities to capture fine details from images. This architecture is famous for being accurate and deep, thus it works well for tasks like e-commerce image classification where products could vary in minor visual differences.

In this application, VGG16 is implemented towards achieving better accuracy concerning the categorization of products based on their visual look. This model has a very useful deep architecture that identifies small details, which include brand logos, specific textures, and small components of the products. Such a model is much applied when the complex pattern is found in images and demands deeper feature extraction.

Thirteen convolutional layers with a small receptive field of 3 x 3 are used to capture high-resolution features from the input image.
Max Pooling Layers: These layers help in reducing the spatial dimensions of the input image after every group of convolution layers; therefore, retain those important features but truncate the complexities.
End.
Fully connected layers: Three fully connected layers follow the convolutional blocks. They produce a combination of all the features extracted and categorize the product image through classification.

Softmax layer: This is an essential output layer applying the softmax function, which gives probabilities for every class and therefore helps in the categorization of the product.



3. **Proposed Model: Hybrid CNN + Grounding DINO**

The Hybrid CNN + Grounding DINO is intended to combine the classic feature extraction technique (CNN) along with the newer method of object detection techniques (Grounding DINO). It blends the strengths of CNN, which correctly identifies pixel-based features such as colors and patterns, with the ability of Grounding DINO to identify semantic features, such as shapes, logos, and text.

.

This proposed model is an enhancement over the usual CNN as well as the VGG16 model since it introduces more diversified feature sets. It can recognize not only the visual features like shape and color but also detect some specific semantic attributes, such as logos, text, or brand markers, in order to precisely classify images related to e-commerce.

K-Means for Color Abstraction: It will use the K-Means pixel values clustering to obtain the most significant color for distinguishing products by appearance.
DINO for Detection: The DINO will attempt to detect special parts in the image to try to recognize logos, shape, or texts of the product.
CNN for Classification: The final CNN is applied so that the color and feature data extracted are combined in order to classify the product image.

Input Preprocessing: Images are preprocessed in order to change colorspace so as to get appropriate space for K-Means Clustering.

K-Means Clustering: This will help in discovering the dominant colors of the images, which will provide a first layer of classification through color.
Grounding DINO: Detects all critical product features such as shape, logo, and text and then generates bounding boxes around these features
CNN: Uses both color and feature-based information to make a final classification of the product image.
Output Layer: It gives as a final output the predicted product category.