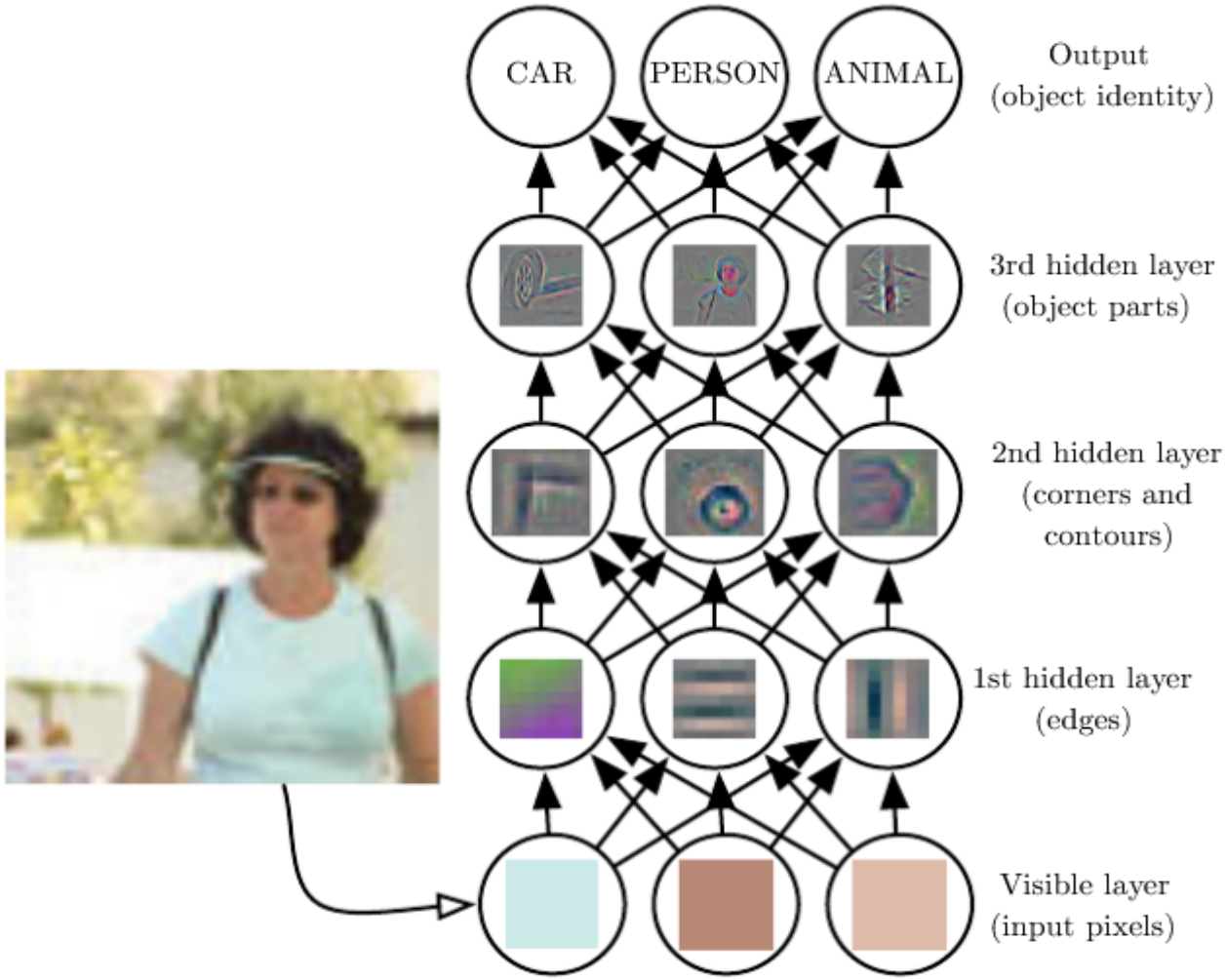
Deep Learning Basics -IT416

Dinesh Naik

Department of Information Technology,
National Institute of Technology Karnataka, India

April 5, 2022

Illustration of a Deep Learning model.



Flowcharts showing how the different parts of an AI system relate to each other within different AI disciplines. Shaded boxes indicate components that are able to learn from data.

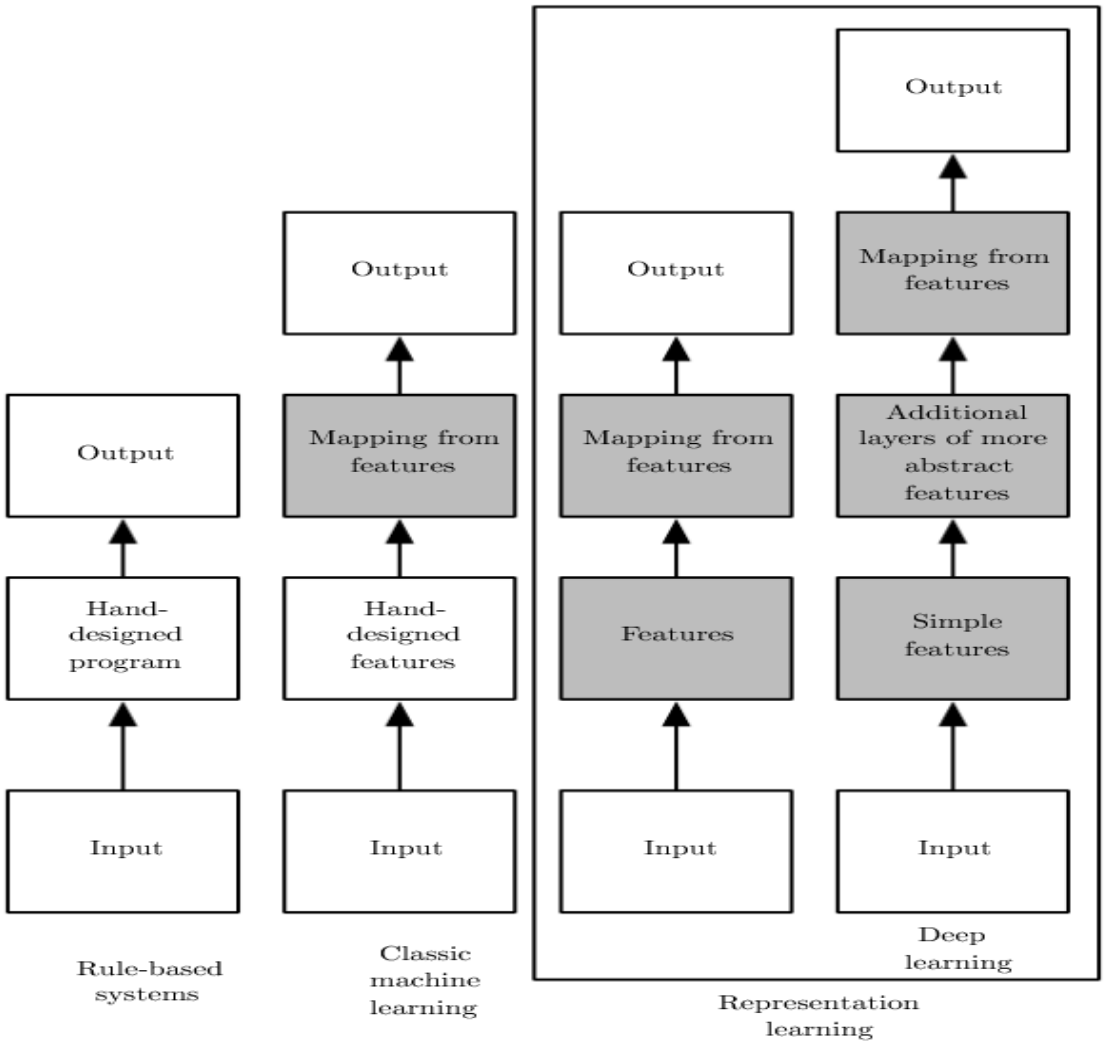


Image and Video Captioning

- Image captioning is a popular research area of artificial intelligence (AI) that deals with image understanding and a language description for that image.
- Image understanding entails detecting and recognizing objects, as well as understanding scene type or location, object properties, and their interactions.
- Generating well-formed sentences requires both syntactic and semantic understanding of the language
- Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories:
- (1) traditional machine-learning-based techniques and
- (2) deep machine-learning-based techniques.

Traditional machine-learning-based

- In traditional machine learning, handcrafted features such as local binary patterns (LBPs), scale-invariant feature transform (SIFT), the histogram of oriented gradients (HOG), and a combination of such features are widely used.
- In these techniques, features are extracted from input data. They are then passed to a classifier such as support vector machines (SVMs) in order to classify an object.
- Since handcrafted features are task specific, extracting features from a large and diverse set of data is not feasible.
- Moreover, real-world data such as images and video are complex and have different semantic interpretations.

Deep machine-learning-based techniques

- Deep machine-learning-based techniques, features are learned automatically from training data and they can handle a large and diverse set of images and videos.
- For example, convolutional neural networks (CNNs) are widely used for feature learning, and a classifier such as Softmax is used for classification.
- CNN is generally followed by recurrent neural networks (RNNs) in order to generate captions.

Image/Video captioning articles into three main categories

- (1) template-based image captioning,
- (2) retrieval-based image captioning, and
- (3) novel image caption generation.
- Most deep-learning-based image captioning methods fall into the category of novel caption generation

Deep-Learning-based image captioning method

- (1) visual space based,
- (2) multimodal space based,
- (3) supervised learning,
- (4) other deep learning,
- (5) dense captioning,
- (6) whole scene based,
- (7) encoder-decoder architecture based,
- (8) compositional architecture based,
- (9) LSTM (Long Short-Term Memory) language model based,
- (10) other language model based,
- (11) attention based,
- (12) semantic concept based,
- (13) stylized captions, and
- (14) novel-object-based.