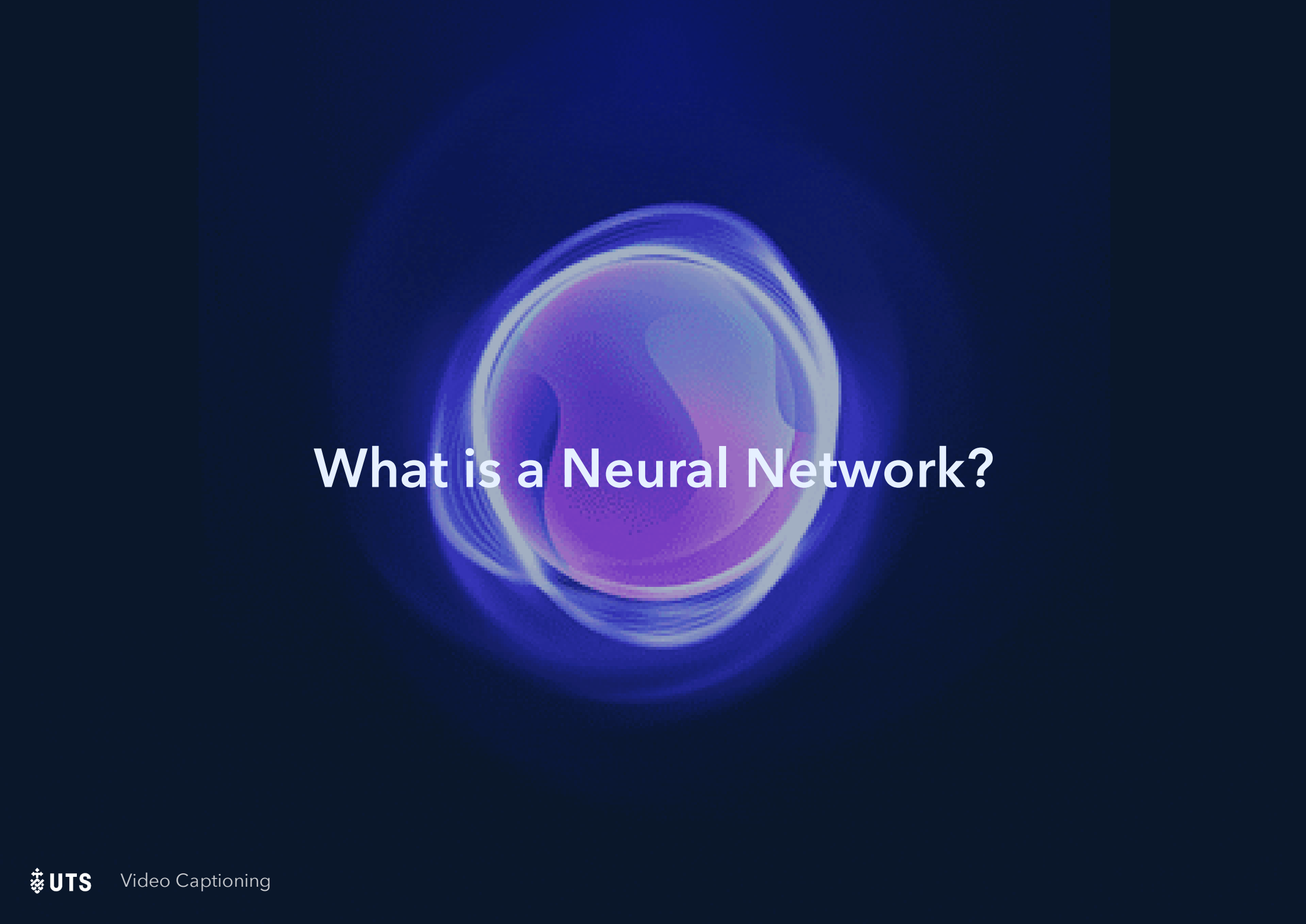
32933 - Research Project

Video Captioning System

Members:
Vraj Mehta - 13488642
Nivetha Anand - 13663024
Prem Rijal - 12957167

Supervisor: Dr Nabin Sharma


What is a Neural Network?

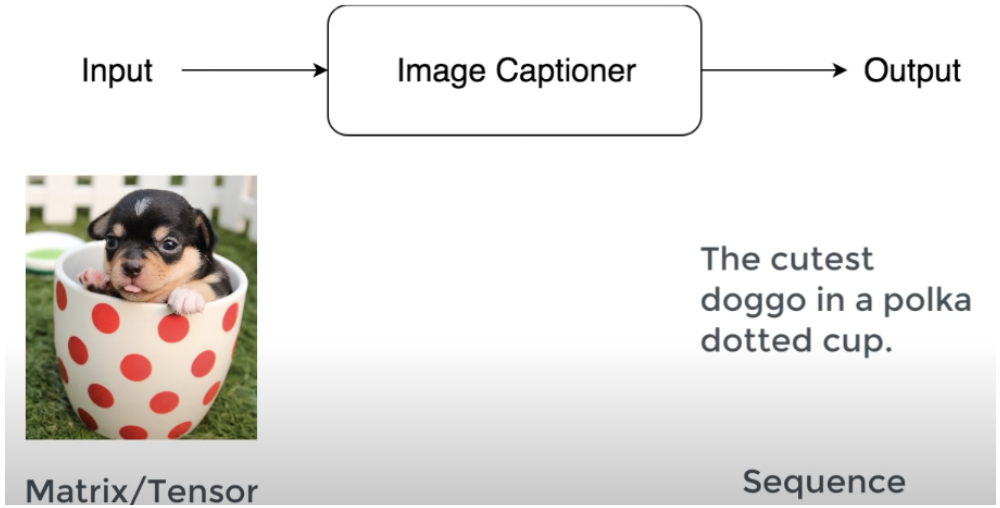
Neural nets are a means of doing machine learning, in which a computer learns to perform some task by analyzing training examples.
Modelled loosely on the human brain, a **neural net** consists of thousands or even millions of simple processing nodes that are densely interconnected.

Various types of Vector-Sequence Problems?

1. Sequence to Vector (Text to Image)

2. Sequence to Sequence (Text to Text)

3. Vector to Sequence (Image to Text)



Vector to Sequence Problem

Applications:

1. Automatic labelling of videos
2. Automated description of images on websites
3. Low vision people who can view larger text fonts
4. Aids in organising photos based on the objects present in the image

Components to Develop Image Captioning Model:

- Dataset
- Architecture
- Neural Network Model
- Evaluation metrics

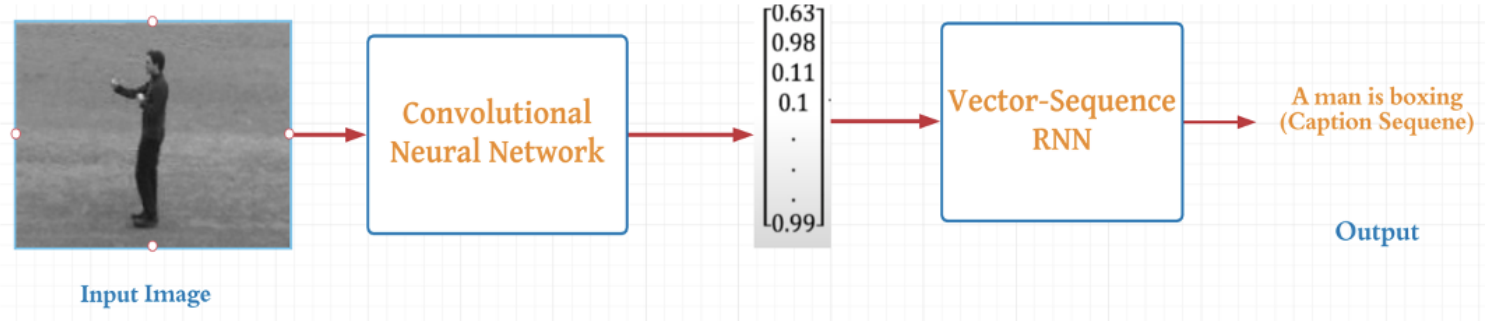
Dataset - KTH Action

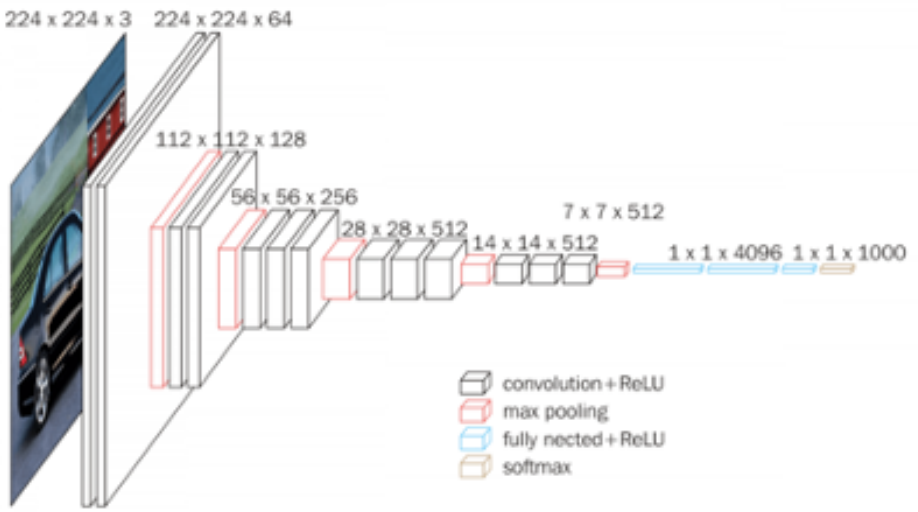
Database Link: https://www.csc.kth.se/cvap/actions/
Total Videos: 600
The dataset is divided into 70% for training, 15% for testing and 15% for training.

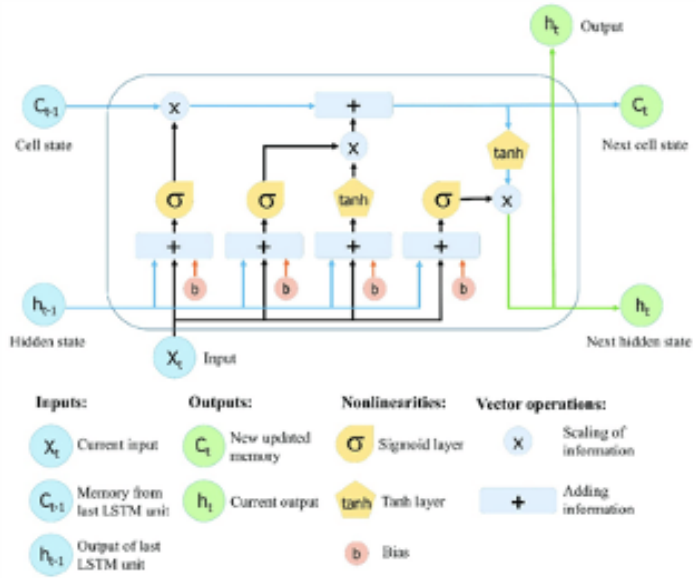
Six actions:
Walking, Jogging, Running, Boxing, Handwaving, Handclapping

System Architecture

- Combination of **CNN** and **RNN**
- **Hybrid** Approach







CNN (Convolutional Neural Network)

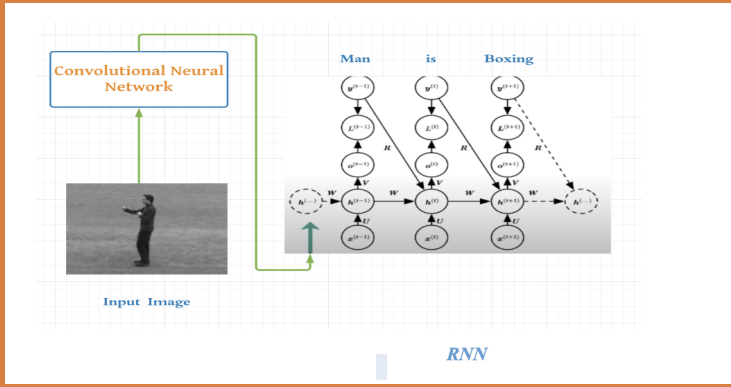
- Used to analyze image data
- Various types layers include Convolutional, Max Pooling, Fully connected
- Applications: Image classification, object detection

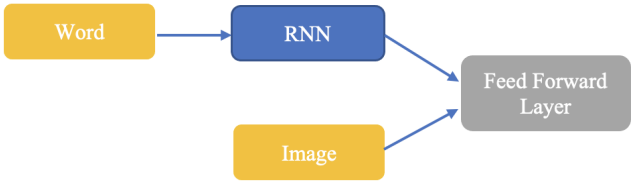
RNN (Recurrent Neural Network)

- Used to analyze sequential data
- Includes LSTM, encoder, decoders
- Applications: Language Modelling and Prediciton

Implementation of Prototype:

a. Preparation of Image Data
b. Preparation of Text Data
c. Development of Deep Learning Model





The **prototype** of our model has been defined by three different sections:

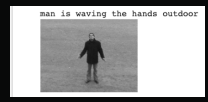
a. Image Feature Extractor: *Pre-trained Model of VGG 16 to extract features*

b. Sequence Processor: *Word Embedding layer to handle text data with LSTM*

c. Decoder: *Combined input from feature extractor and sequence processor is processed by a dense layer to make final predictions*

Results and Evaluation

- Trained for 20 epochs
- Training Loss: **0.3464**, Validation Loss: **0.3673**



Metrics Used: BLEU **(**Bilingual Evaluation Understudy)

1. BLEU-1: 0.357752
2. BLEU-2: 0.598124
3. BLEU-3: 0.734640
4. BLEU-4: 0.773384

Conclusion:

- ☑ Created Dataset (Referred from KTH Action)
- ☑ Developed System Architecture
- ☑ Trained and Validated the Model
- ☑ Tested against Evaluation Metrics

Future work:

1. Develop a REST API using Python/Django framework
2. Design a frontend web-app or mobile-app
3. Investigate "***Attention Mechanism"*** for further improvements


T'S A WRAP