CNN-Based Video Decoding and Analysis

Bikkavolu Prasanthi (B20CS010)

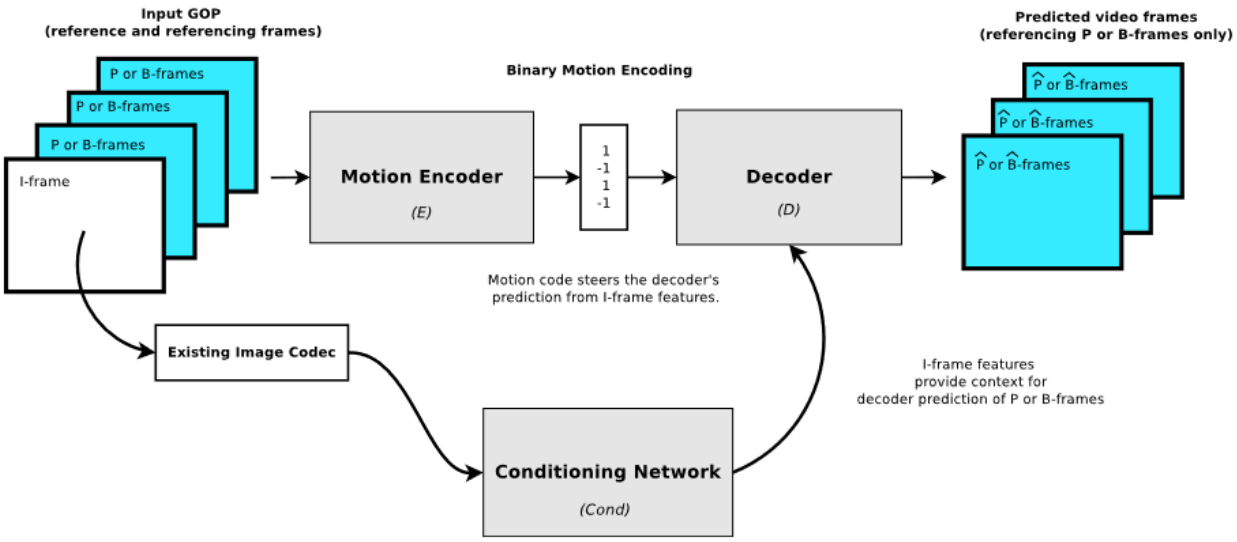
INTRODUCTION

By 2022, it is expected that 82% of Internet traffic will be video. Mathematical algorithms are used in conventional video decoding techniques to decode compressed visual data.In order to reduce bandwidth usage and large bitrates, video compression must be improved.In our method, we conduct video encoding and decoding tasks using a CNN-based network, which is essentially a neural network. It is built on a deep neural network design that learns to encode and decode video data in one step, eliminating the need for additional encoding and decoding steps.

METHODOLOGY

Overview

The neural network **encoder E** compresses and binarises the motion occurring in a Group Of Pictures (GOP), a video segment containing designated reference (I) and referencing (P) frames. At the **decoder D**, I-frame features extracted by the conditioning network '**Cond**' are transformed based on the information held in the binarised motion encoding to predict the P referencing frames. Note that 'Cond' is not responsible for I-frame compression, this is done by an existing image codec.



Dataset

By using the Vimeo-90k dataset, which was recently developed for testing various video processing tasks like video denoising and video super-resolution, we train the suggested video compression framework. It comprises 89,800 separate clips with distinctive material that are not connected to one another. We test our suggested algorithm on the UVG dataset and publish the results to show the effectiveness of our suggested approach. These datasets have a variety of contents and resolutions and are frequently used to gauge how well video compression methods work.

Motion Vectors

In order to depict how pixels move from one frame to the next, motion vectors (MVs) are utilized. Dense optical flow, however, generates too many MVs for effective compression. As a result, block-based motion estimation and compensating algorithms are used by video codecs. Using an MV that contains its displacement in the x and y dimensions, each macroblock in the current frame is connected to the position of the most comparable macroblock in a previous or subsequent reference frame in this procedure. The search for representative macroblocks in the reference frame is accelerated using a variety of search techniques. Only MVs are required to be transmitted following the transmission of a reference I-frame in order to motion-compensate macroblocks in the I-frame and create predictions for the succeeding frames in a video sequence. To enhance the quality of the reconstruction, the residuals (differences) between the vector-based motion predictions and the original video frames are encoded using standard image compression.

Encoder

The motion vectors are then used as input to the **CNN**, which is responsible for encoding the video frame into a compressed format. The CNN typically consists of multiple convolutional layers, which are used to extract features from the input data, and pooling layers, which reduce the size of the feature maps.Once the features have been extracted and pooled, they are typically passed through fully connected layers, which perform the actual encoding of the video frame. The output of the encoder is a compressed representation of the video frame, which can be stored or transmitted more efficiently than the original uncompressed video.

Conditioning Network

With the help of previously encoded frames, the conditional network in video codecs makes predictions about next frames in a video sequence. It requires a collection of reference frames as

input, often one or more completely encoded I-frames, which are used to condition the prediction of upcoming frames. With the intention of minimizing the discrepancy between the anticipated frames and the actual frames in the video sequence, the network is trained using a supervised learning methodology. The conditional network classifies the frames during encoding and compares them to the real frames; the difference is then sent to the decoder.

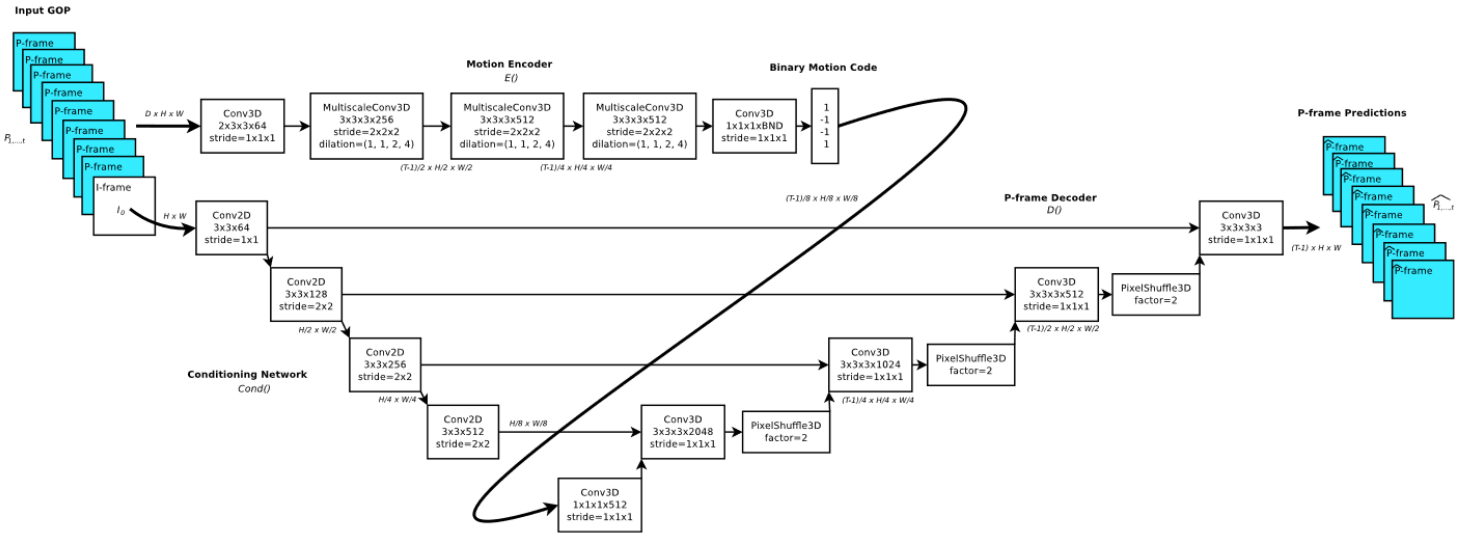
Decoder

Future frames in the video sequence are predicted by the **decoder** using the motion encoding. The original I-frame material is available to the decoder during training, and during test time, an existing picture codec separately encodes and decodes the I-frames. The bitrate is controlled by the number of output channels in the final encoder layer, and the encoder always applies an 8x compression to the input GOP's width, height, and time axis. Depending on whether it is conditioned on a single I-frame or a pair of bounding I-frames, the decoder either conducts motion-guided extrapolation or interpolation.

Overall, binarized motion encoding and conditioned decoding increase compression and reconstruction quality, and the codec's design is completely convolutional to support a wide variety of input frame-sizes and dynamic GOP durations.

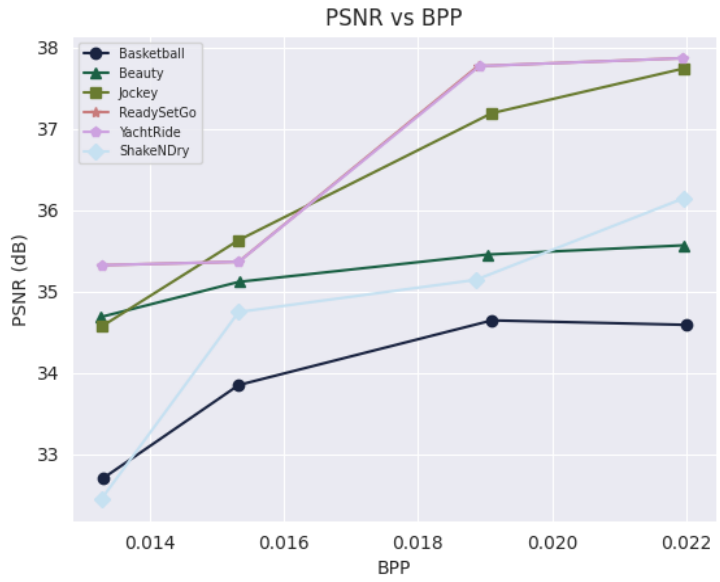
During training we use a L2 reconstruction loss:

$$LR = ||P - P\hat{}||2$$



RESULTS

Peak Signal to Noise Ratio (PSNR), an objective evaluation metric, is used to rate the quality of our predicted video frames.We use bits per pixel (Bpp) to represent the required bits for each pixel in the current frame in order to calculate the amount of bits needed to encode the representations.

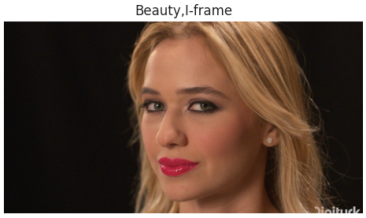


Basketball,I-frame

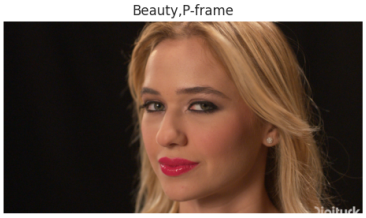
Basketball,P-frame

Basketball,Predicted P-frame PSNR : 34.58821629888925

Beauty,I-frame



Beauty,P-frame

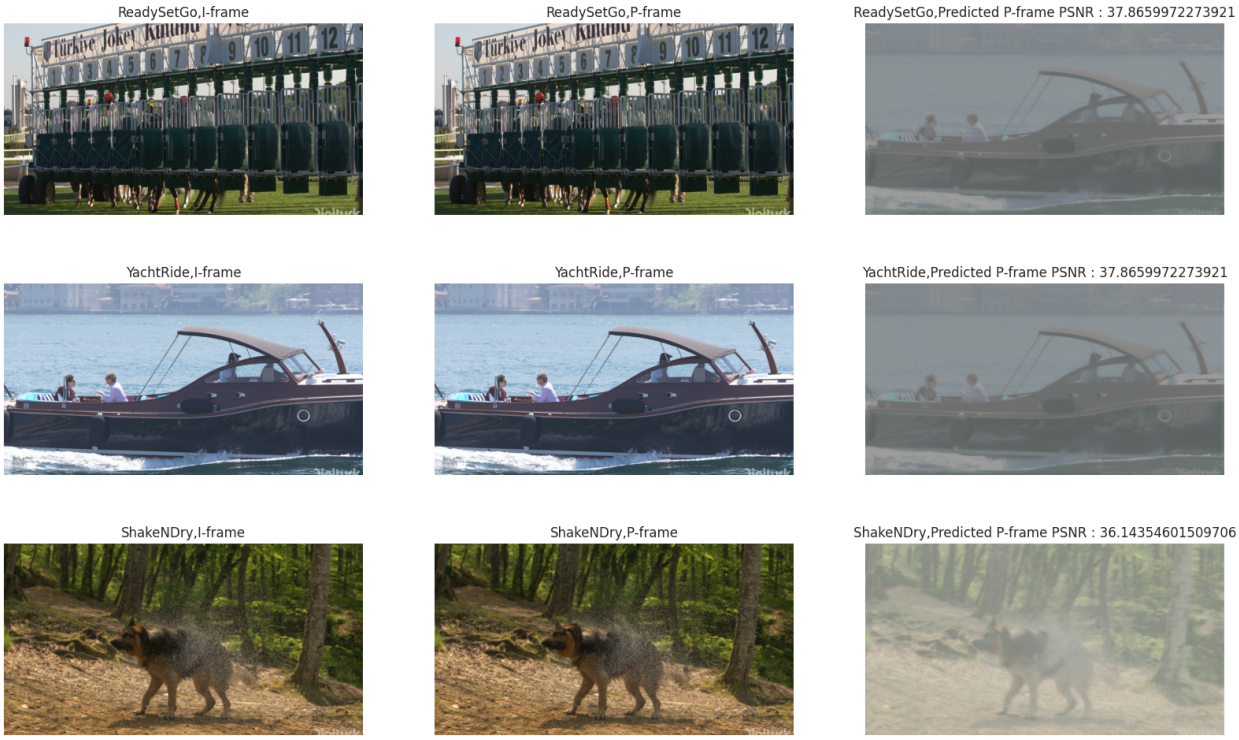


Beauty,Predicted P-frame PSNR : 35.587945936316046

Jockey,I-frame

Jockey,P-frame

Jockey,Predicted P-frame PSNR : 37.73934095782754



Figures show the visual quality of the reconstructed frames for different videos . Leftmost is the reference frame. Middle is the original frame. Rightmost is the reconstructed frame.

REFERENCES

1. Dong Liu, Yue Li, Jianping Lin, Houqiang Li, Feng Wu ; Deep Learning-Based Video Coding: A Review and A Case Study

2. Guo Lu1, Wanli Ouyang2, Dong Xu3, Xiaoyun Zhang1, Chunlei Cai1, and Zhiyong Gao ; DVC: An End-to-end Deep Video Compression Framework

3. Andr´e Nortje, Herman A. Engelbrecht, Herman Kamper ; Deep motion estimation for parallel inter-frame prediction in video compression