Quantitative Estimation of Video Pixel Quality For Action Recognition

Submitted by: Mansi Manoj Maurya

Introduction

The primary goal of this project is to develop models that can automatically **predict video quality** based on specific **spatial features** such as sharpness, brightness, contrast, noise, and blur.

The objective is to assign each video a **quality score between 0 and 1**, where 0 represents poor quality and 1 represents high quality.

The motivation behind this task is the increasing need to assess and filter large volumes of video data, particularly when video quality can affect the performance of downstream tasks like content analysis, video streaming, or action recognition.

Dataset Overview

Dataset Structure:

- Videos were artificially degraded to simulate various quality levels.

Degradations Applied:

- **Brightness**: Adjusted to simulate overexposure or underexposure.
- **Contrast**: Modified to simulate high and low contrast levels.
- **Sharpness**: Reduced using Gaussian blur to introduce varying degrees of blurriness.
- **Noise**: Added to simulate video artifacts.

Ground-Truth Scores:

- Each video was assigned a **quality score** ranging from **0** (worst) to **1** (best).
- The dataset includes original, high-quality videos and multiple degraded versions.

Methods

Extracted Features manually.

Multi-Layer Perceptron (MLP):

- Uses **hand-crafted spatial features** (sharpness, brightness, contrast, noise).
- **Input**: Feature vector representing the video.
- **Architecture**: Fully connected layers for predicting video quality scores.
- **Output**: Quality score between 0 (worst) and 1 (best).

Temporal Segment Network (TSN):

- Samples multiple frames from each video.
- **ResNet** backbone extracts deep spatial features.
- **Aggregates** features from the sampled frames.
- **Output**: Video quality score prediction, leveraging features from multiple frames.

Results

MLP Results:

- **R^2 score** of 0.54 (explains 54% variance).
- Reasonable performance in predicting video quality, relying on hand-crafted spatial features.

TSN Results:

- **TSN significantly outperformed MLP**:
  - Higher **R^2 score**, showing stronger predictive power.
  - Better handling of multiple frames through ResNet's deep feature extraction.
  - Improved accuracy and consistency across various degraded video conditions, leveraging spatial features from multiple frames.
- **Overall**: TSN's ability to capture richer spatiotemporal features led to a more accurate video quality prediction compared to MLP.

Conclusion

- Successfully developed models to predict video quality based on spatial features.
- MLP is effective for basic video quality predictions using hand-crafted features.
- **TSN is the more effective model** for video quality prediction due to its ability to capture rich, multi-frame spatial features.