NTIRE 2025 Short-form UGC Video Quality Assessment and Enhancement Challenge-Track1-VQA

[Nourayn]

Team Name

Nourayne

Team Members

- **Team Leader**: Nourine Mohammed Nadir (nounadir@gmail.com)
- **Member 1**: Nourine Mohammed Nadir

Method Overview

Our solution for the NTIRE2025 Video Quality Assessment Challenge is a **two-stage deep learning model** that combines **spatial feature extraction** using a pre-trained ResNet-50 backbone and Faster-RNN with **temporal modeling** using a Bidirectional LSTM (BiLSTM). The model is designed to predict the **Mean Opinion Score (MOS)** for video quality assessment by capturing both spatial and temporal information from videos.

Architecture

Feature Extraction

- **Backbone**: We use a **ResNet-50** model pre-trained on ImageNet to extract spatial features from video frames.
- **Region Proposal Network (RPN)**: A Faster R-CNN is employed to generate region proposals and extract region-of-interest (RoI) features.
- **Output**: For each video, we extract a fixed number of frames (`num_frames`), and each frame is processed to produce a **2048-dimensional feature vector**.

Temporal Modeling

- **BiLSTM**: The extracted features from all frames of a video are passed through a **Bidirectional LSTM** to capture temporal dependencies.
- **Fully Connected Layers**: The output of the BiLSTM is fed into a series of fully connected layers with **ReLU activations**, **dropout**, and **Layer Normalization** to prevent overfitting and stabilize training.
- **Output Layer**: A single regression head predicts the MOS score for the video.

Loss Function

The model is trained using a **composite loss function** that combines:

- **Mean Squared Error (MSE)**: To minimize the difference between predicted and ground truth MOS scores.
- **Ranking Loss**: To ensure that the model correctly ranks videos based on their quality.

The final loss is a weighted sum of MSE and Ranking Loss:

$$\text{Loss} = 0.8 \cdot \text{MSE} + 0.2 \cdot \text{Ranking Loss}$$

Training Pipeline

Feature Extraction

- Videos are preprocessed to extract a fixed number of frames.
- ResNet-50 and Faster R-CNN are used to extract spatial and RoI features.
- Features are saved for training and evaluation.

Training

- The BiLSTM model is trained on the extracted features using the composite loss function.
- Training is performed on a combination of training and validation datasets to improve generalization.

Testing

- The trained model is evaluated on the test dataset.
- Metrics such as **SROCC**, **PLCC**, **KROCC**, and **RMSE** are computed to assess performance.

Key Features

- **Efficient Feature Extraction**: Leveraging pre-trained ResNet-50 and Faster R-CNN ensures robust spatial feature extraction.
- **Temporal Modeling**: BiLSTM captures temporal dynamics in videos, which is critical for VQA.
- **Composite Loss**: Combines regression and ranking objectives to improve prediction accuracy and ranking consistency.
- **Modular Design**: The pipeline is modular, allowing easy integration of new datasets or models.

Results

The model achieves competitive performance on the challenge dataset, with strong correlation metrics (SROCC, PLCC) and low error rates (RMSE). Detailed results are provided in the submission.

Usage

To reproduce the results:

- Run `main.py` for feature extraction, training, and testing.
- Adjust hyperparameters (e.g., `num_frames`, `batch_size`, `learning_rate`) as needed.
- Refer to the `README.md` for detailed instructions.

Model pipeline