ISSN No:-2456-2165

Video Quality Assessment (VQA) using Vision Transformers

Kallam Lalithendar Reddy; Pogaku Sahnaya; Vattikuti Hareen Sai; Gummuluri Venkata Keerthana
Department of Computer Science and Engineering
VNR Vignana Jyothi Institute of Engineering and Technology, India

Abstract:- In this paper, we check the potential of vision transformers in the field of Video Quality Assessment (VQA). Vision Transformers (ViT) are used in field of computer vision based on working nature of transformers in Natural Language Processing (NLP) tasks. They work on the relationship between the input tokens internally. In NLP we use words as tokens, whereas in computer vision we use image patches as tokens where we try to capture the connection between different portions of the image. A pre-trained model of ViT B/16 over imageNet-1k was used to extract features from the video and to validate them over the MOS scores of the video. The patch embeddings are given tokens called as positional embeddings and are send to transformer encoder. There are total 12 layers in ViT - Base Transformer Encoder. Each encoder has a Layer Norm, Multi-Head Attention followed by an another Layer Norm with Multi-Layer Perceptron (MLP) block. The classifier head of the Transformer was removed to get feature vector as our aim is not to classification. After the features are achieved we use an Support Vector Regressor (SVR) of Radial Basis Function (RBF) kernel to assess the video quality.

Keywords:- *Konvid 1-k Dataset, Vision Transformer, Support Vector Regressor, Attention, Token Embeddings.*

I. INTRODUCTION

Before the origin of Vision Transformers, we used Convolutional Neural Networks (CNN) models for image related tasks for years. Through the use of filters these CNNs were able create feature maps of the input image and highlight the most relevant parts of the image and send them to Multi-Layer Perceptron for further tasks. The advantage of using CNN architecture is that they avoid the need for hand designed visual features. Instead learn directly from data end to end. The CNN architecture itself on a whole is designed for images. As everyone wants a single model which performs multiple tasks. Then researchers thought of building architecture which were more leverage more domain agnostic and computationally efficient and fast.

In this process, Transformers were designed. Transformers are first introduced in the paper "Attention is All You Need". At the beginning, they are used for text-based tasks as they use the concept of Attention Mechanism. Transformers have very high success rates in NLP tasks as they enable long term dependencies between the sequence of input elements. Transformers initially found their applications in Natural Language Processing by models like BERT and GPT-3. The design of transformers now works with many fields like images, videos, text, speech etc.

Attention Mechanism plays a prominent role in transformers. It enhances the important field of the input and fades out the rest. CNN does not encode the relative position of different features. Large Receptive Fields are required in order to track long-term or long-range dependencies within the input data. This problem is overthrown with the use of transformers with attention. Self-attention also called as Intra attention which allows every element of a sequence to interact with every other element and find out to which we need to pay more attention to. It shows the long-term dependencies in the sequence of data. Self-attention is a weighted combination of the embeddings. The one which is more closely dependent or related is given more weight-age in the sequence.

The original transformers take text as input tokens and then uses it for classification, translation and other NLP related tasks. Few modifications are made to transformer to design ViT to make it work on images and know how much it learns from the image structure. ViT divides the image into patches according to its kernel and stride size. Each patch is flattened into single vector by concatenating all the channels of image like for example, RGB channels of the image. As transformers are agnostic to its structure, we add positional embeddings to the patches of the image to learn more about the relationships, hierarchies and alignments between the patches and structure of the images.

In its first attempt ViT was trained on imageNet, they achieved 77.9 whereas CNN the showed an accuracy of 85.8. Then, studies made understand that ViT is overfitting the data due to lack of knowledge on the image related data. After that ViT was trained on imageNet-21k (14M images with 21k classes) and JFT (300M images with 18k classes) and is compared over existing state-of-art CNN model. The ViT totally outperforms the CNN. ViT with sufficient data has an excellent performance. So, ViT are trained over 600M parameters and then results are observed. We can see that large ViT attains a benchmark of 88.55 on imageNet and 99.50 on CIFAR-10. ViT outperforms state-of-art CNN by 4 four times in terms of computational efficiency and accuracy.

We now take a ViT B/16 pre-trained on imageNet-1k where 16 is the patch size has 12 layers with hidden size (Dimensionality of encoder layers and pooler layers) of 768 and 12 attention heads. We train it over the Konvid-1k data set and extract the features of the videos, flatten the matrix and make feature vectors for the data set. These features are

used by Support Vector Regressor (SVR) along with their MOS scores to predict the score of the video in the range of 1-5. The results shown that vision transformer performed better than I3D with Resnet50 pretrained with Kinetics 400.

II. RELATED WORK

- **Introduced by Dosovitskiy et al. in An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale:** The Vision Transformer, or ViT, is a model for image classification that employs a Transformer-like architecture over patches of the image. An image is split into fixed-size patches, each of them are then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder. In order to perform classification, the standard approach of adding an extra learnable "classification token" to the sequence is used.

- **Introduced by Yuan et al. in Tokens-to-Token ViT: Training Vision Transformers from Scratch on Image Net:** T2T-ViT (Tokens-To-Token Vision Transformer) is a type of Vision Transformer which incorporates 1) a layer wise Tokens-to-Token (T2T) transformation to progressively structure the image to tokens by recursively aggregating neighboring Tokens into one Token (Tokens-to-Token), such that local structure represented by surrounding tokens can be modeled and tokens length can be reduced; 2) an efficient backbone with a deep-narrow structure for vision transformer motivated by CNN architecture design after empirical study.

- **Introduced by Ranftl et al. in Vision Transformers for Dense Prediction Dense**: Prediction Transformers (DPT) are a type of vision transformer for dense prediction tasks. The input image is transformed into tokens (orange) either by extracting non-overlapping patches followed by a linear projection of their flattened representation (DPT-Base and DPT-Large) or by applying a ResNet-50 feature extractor (DPT-Hybrid). The image embedding is augmented with a positional embedding and a patch-independent readout token (red) is added. The tokens are passed through multi- ple transformer stages. The tokens are reassembled from different stages into an image-like representation at multiple resolutions (green). Fusion modules (purple) progressively fuse and up sample the representations to generate a fine-grained prediction.

- **Introduced by Han et al. in Transformer in Transformer:** Transformer is a type of self-attention-based neural networks originally applied for NLP tasks. Recently, pure transformer-based models are proposed to solve computer vision problems. These visual transformers usually view an image as a sequence of patches while they ignore the intrinsic structure information inside each patch. In this paper, we propose a novel Transformer-in-Transformer (TNT) model for modeling both patch-level and pixel-level representation. In each TNT block, an outer transformer block is utilized to process patch embeddings, and an inner transformer block extracts local features from pixel embeddings. The pixel-level feature is projected to the space of patch embedding by a linear transformation layer and then added into the patch. By stacking the TNT blocks, we build the TNT model for image recognition.

- **Introduced by Jiang et al. in All Tokens Matter: Token Labeling for Training Better Vision Transformers** LV-ViT is a type of vision transformer that uses token labelling as a training objective. Different from the standard training objective of ViTs that computes the classification loss on an additional trainable class token, token labelling takes advantage of all the image patch tokens to compute the training loss in a dense manner. Specifically, token labeling reformulates the image classification problem into multiple token-level recognition problems and assigns each patch token with an individual location-specific supervision generated by a machine annotator.

III. METHODOLOGY

A. Data Set

The goal of Video Quality Assessment (VQA) is to find the quality of digital videos. Video Quality Assessment (VQA) strongly depends on semantics, context and types of visual distortions. We used KonViD-1k Database which consists of large no. of natural, real- world video sequences with their corresponding Subjective Mean Opinion Scores (MOS). The data set contains total of 1200 videos. While each video has an individual ID named as flickr ID which helps to find video on Flickr.com. Each file is a 8 seconds video and is given a quality score from 1 to 5. Fig 2. shows the no. of videos in the dataset for a particular MOS range. Every video has varying frame rate.

ISSN No:-2456-2165

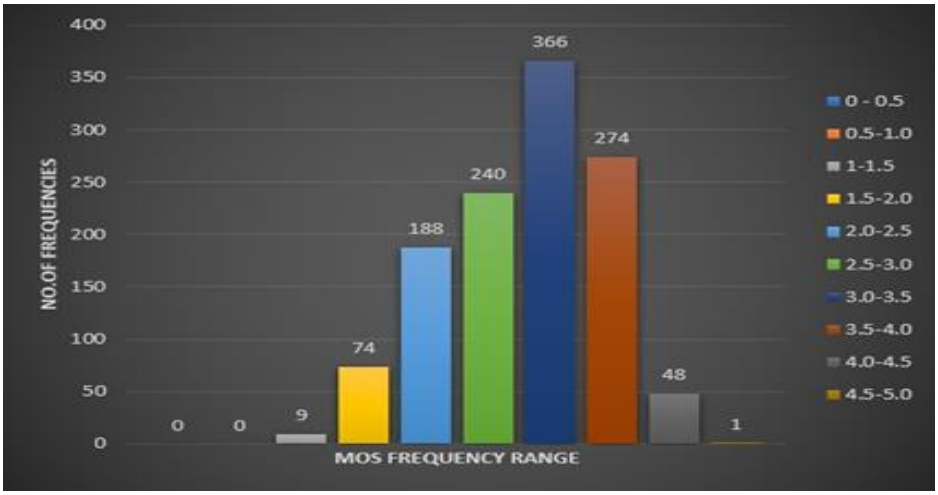

Fig. 2: Data Visualization

B. Data Cleaning Phase

As mentioned, each video has a duration of 8 seconds. The frame width and frame height of videos are 960 x 540 pixels. ViT expects each image sent to be of same resolution. The ViT is pre-trained on a resolution of 224 x 224. During fine-tuning of the model, it is beneficial to use an image of higher resolution than pre-training because there may be significant discrepancy between the structures identified in the image during the train and test time. So, all the videos frame size is adjusted to 384 x 384 pixels. The varying frame rate also becomes a problem while extracting frames from the video as a part of it are 23.3 fps and some are 29.97 fps. A good frame rate is between 24-30 frames per second. Any- thing less than 20 frames-per-second will result in choppy motion in the video. Increasing the frame rate beyond the original frame rate will not produce smoother video quality. It will just result in duplicate frames and a larger video file. So, we convert all the videos into 30 frames-per-second.

C. Feature Extraction Using Vision Transformer

Vision Transformer (ViT) is introduced in a research paper published as a conference paper at ICLR 2021 titled "An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale".

The following steps are processed for feature extraction of a video:

- Split each video of database into 25 frames.
- Send each frame through an already pre-trained ViT B/16 model trained on ImageNet-1k. Obtain the Feature vector for the frame.
- Obtain feature vector for all 25 frames of the video.
- Now flatten feature matrix into single 1D vector or convert into 1D vector by taking mean of all 25 frames feature vectors.
- Using this technique, extract feature vectors for all videos using ViT.

➢ *Splitting of video into frames*

Video can't be studied on a whole directly. We need to divide the video into frames to study the video. Hence, we divide it into 25 frames to study the motion of the tape.

➢ *Obtaining Feature Vector for each frame*

The resolution of each frame is 384 x 384. The ViT B/16 has kernel size of 16 x 16. So, each frame is divided into 24 x 24 patches. In Fig 3 We can see how the image is divided into patches. Next the ViT passes them through linear projection layer where we get 1 x 768 projection vector for each patch. We have total of 24 x 24 i.e, 576 patches. In Fig 4, we can observe the process of how the image patches of image are processed. So, we get 576 x 768 matrix. Class tokens are added to patch embedding matrix now it becomes 577 x 768 along with the positional embeddings and it next sent to transformer encoder. The Fig 5 shows the layers residing in transformer encoder. The positional embeddings are given to know the relation between the information of patches. Transformer encoder consists of two blocks called Multi-head Attention and Multi-Layer Perceptron (MLP) blocks. The combined matrix is sent as input Layer Norm and then sequentially to first layer of encoder where it undergoes qkv (queries, keys, values) attention. The input matrix is converted into 577 x 2304(768 x 3) where each attention matrix has a shape of 577 x 768. The reshaped form of matrix 12 x 577 x64 shows that it passes through 12 attention heads. SoftMax function is used to get attention matrix and it is linearly projected to get 768 features and again passes through the Layer Norm before sending it to MLP unit. Finally, after the 12th layer we get feature vector of size 1 x 768 i.e., it returns no. of hidden features.

ISSN No:-2456-2165

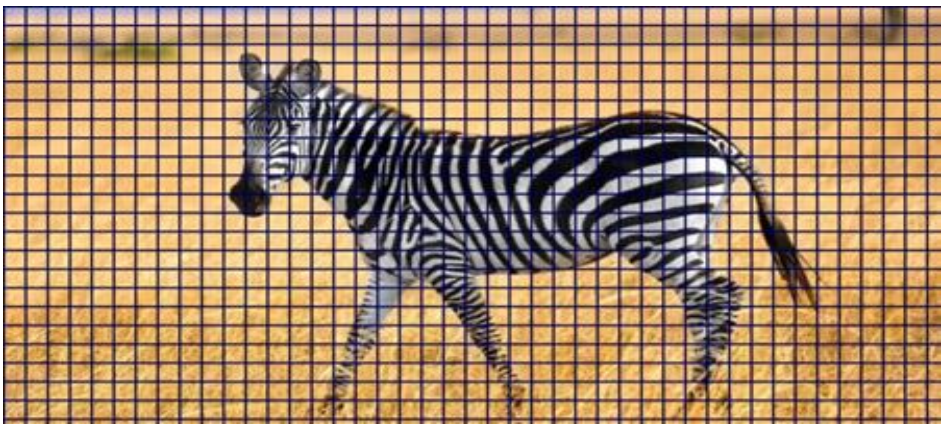

Fig. 3: Splitting up of a frame into patches.

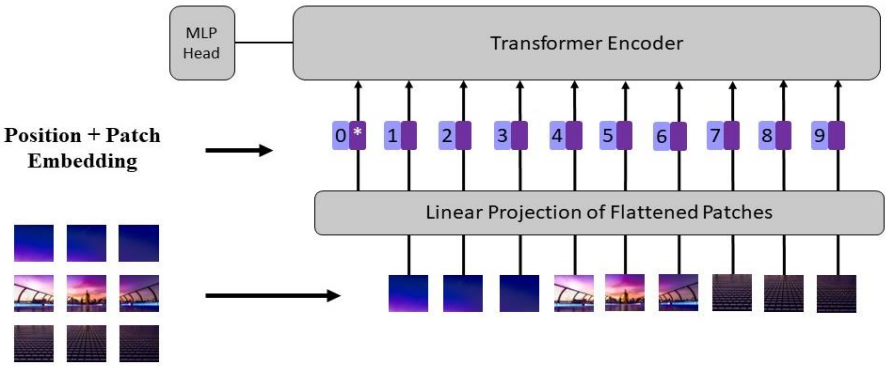

Fig. 4: Vision Transformer (ViT)

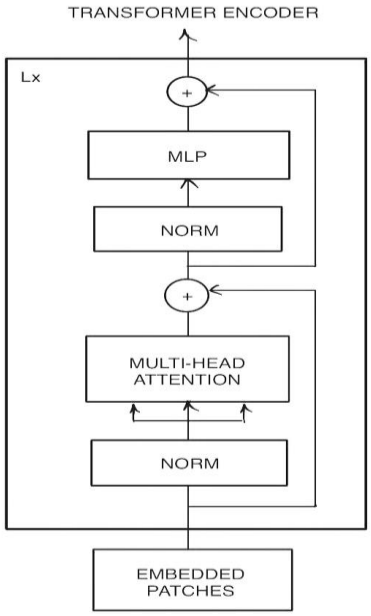

Fig. 5: Transformer Encoder

➢ *Obtain features of all frames*

We get 1 x 768 feature vector for each frame. As we have 25 frames for each video, we get a feature matrix of size 25 x 768.

➢ *Flattening of feature matrix*

Each video now has a feature matrix of 25 x 768. We flatten the feature matrix by taking the mean of all the rows of matrix and convert it into a single feature vector of size 1 x 768.

➢ *Extract feature vectors for all videos*

Now, we know the process of how to extract the features of video. Using the same procedure draw out features of remaining videos in the data set.

IV. APPLY SUPPORT VECTOR REGRESSION MODEL

After extracting the features from all the videos, we get a data set consisting of 1200 rows and 768 columns. Attach MOS scores of corresponding videos to its feature vector. Now split the data into train and test data. Fit Support Vector Regression model on training set and specify the 'rbf' kernel. Now, predict the MOS scores of test data set.

V. RESULTS

Initially, first we fit the data set building a Dummy regression model. This Dummy Regressor makes predictions with the use of easy policies. We used mean strategy in this model. This regressor is beneficial as a simple baseline to evaluate with different (actual) regressors.

Table 1: Dummy Regressor

Metrics	Scores
Mean Square Error	0.5180
Mean Absolute Error	0.557
Root Mean Square Error	0.7197
R-squared value	0.0008

The next study is performed building a Linear Regression model. It gives mathematical approach about the relation between two or more independent variables and a dependent variable. It can be linear or non-linear.

Table 2: Linear Regression

Metrics	Scores
Mean Square Error	0.3157
Mean Absolute Error	0.3693
Root Mean Square Error	0.5619
R-squared value	0.3900

The next evaluation is done using Support Vector Regression (SVR). It is same as SVM but SVR is used to find the best fit line.Here, we have used SVR with 'rbf' kernel and the other parameters are set to their default values.

Table 3: Support Vector Regressor

Metrics	Scores
Mean Square Error	0.3067
Mean Absolute Error	0.4369
Root Mean Square Error	0.5538
R-squared value	0.4074

Video quality assessment (VQA) methods quantify the quality of a video. The results of proposed approach were compared with results using other architectures and with the subjective mean opinion scores (MOS). And the experimental results expressed that the proposed metric method using vision transformers could more accurately measure the mean opinion scores. The Pearson Correlation Coefficient reaches a high value of 0.709 and Spearman Correlation Coefficient achieves 0.59.

Table 4: Co-Relation Coefficients

Model Used	Pearson	Spearman
Pre-trained ViT B/16	0.709	0.59
I3D with Resnet50	0.59	0.57

VI. CONCLUSION

we worked on video quality assessment (VQA) for video converters (ViT) that provide video features. We validated them with video MOS points, pre-trained ViT B/16 model with imageNet-1k. We took the KonViD-1k database, which consists of real video episodes and their corresponding subjective mean opinion scores (MOS). The material contains a total of 1200 videos. All video frames are then set to 384 x 384 pixels. Then we converted all videos to 30 fps. After extracting features from all videos, we concatenated the MOS scores of the corresponding videos into its feature vector and divided the data into train and test data. A Fit Support Vector Regression model is applied to the training set and the "rbf" kernel was determined, and then the MOS scores of the test dataset were predicted. Since classification is not our goal, the Transformer classification head has been removed to obtain the feature vector. After obtaining the features, we evaluated the video quality using Support Vector Regressor (SVR) and Radial Basis Function (RBF) kernel. This allows us to perfectly collect global contexts. This can be very important to understand the entire visual content and identify potential multi-frame quality issues when evaluating video quality.

REFERENCES

[1]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[2]. Alammar, J (2018). The Illustrated Transform. Retrieved from https://jalammar.github.io/illustrated-transformer/
[3]. Alammar, J (2018). The Illustrated Transformer. Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention). Retrieved from https://jalammar.github.io/illustrated-transformer/
[4]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[5]. Kancharla, P., & Channappayya, S. S. (2021). Completely blind quality assessment of user generated video content. IEEE Transactions on Image Processing, 31, 263-274.

[6]. Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In International conference on machine learning (pp. 7354-7363). PMLR.
[7]. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., & Rabbat, M. (2021). Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8443-8452).
[8]. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).
[9]. Götz-Hahn, F., Hosu, V., Lin, H., & Saupe, D. (2021). KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. IEEE Access, 9, 72139-72160.
[10]. Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., ... & Saupe, D. (2017, May). The Konstanz natural video database (KoNViD-1k). In 2017 Ninth international conference on quality of multimedia experience (QoMEX) (pp. 1-6). IEEE.
[11]. Chetouani, A. (2020, July). Image quality assessment without reference by mixing deep learning-based features. In 2020 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
[12]. Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., & Beslay, L. (2019, June). Faceqnet: Quality assessment for face recognition based on deep learning. In 2019 International Conference on Biometrics (ICB) (pp. 1-8). IEEE.
[13]. Amirshahi, S. A., Pedersen, M., & Yu, S. X. (2016). Image quality assessment by comparing CNN features between images.
[14]. Bosse, S., Maniry, D., Wiegand, T., & Samek, W. (2016, September). A deep neural network for image quality assessment. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 3773-3777). IEEE.
[15]. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., ... & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 558-567).
[16]. Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12179-12188).
[17]. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. Advances in Neural Information Processing Systems, 34, 15908-15919.
[18]. Jiang, Z. H., Hou, Q., Yuan, L., Zhou, D., Shi, Y., Jin, X., ... & Feng, J. (2021). All tokens matter: Token labeling for training better vision transformers. Advances in neural information processing systems, 34, 18590-18602.
[19]. Touvron, H., Vedaldi, A., Douze, M., & Jégou, H. (2019). Fixing the train-test resolution discrepancy. Advances in neural information processing systems, 32.

Fig. 1: Splitting of a video into frames