DCVQE: A Hierarchical Transformer for Video Quality Assessment (Supplementary Material)

Zutong Li and Lei Yang

Weibo R&D Limited, USA
{zutongli0805, trilithy}@gmail.com

1 Selection of the CNN Backbone

We test two series of architectures based on ResNet and ViT with different settings. As Table 1 shows, we determine that deeper structure results in worse system performance, and partially fine-tuned backbones usually work better than fully fine-tuned ones. For example, the fully fine-tuned ResNet-18$_{full}$ outperforms its corresponding deeper versions ResNet-50$_{full}$ and ResNet-101$_{full}$, while the partially fine-tuned ResNet-50$_{partial}$ shows a significant improvement over its fully fine-tuned one ResNet-50$_{full}$. The same conclusion can be drawn by analyzing the ViT results. Also, we incorporate an attention-based IQA architecture PHIQNet as the feature extractor to our model. Table 2 shows the performance comparisons of two models (with different backbones). As seen, the partially fine-tuned ResNet-50 and PHIQNet contribute similarly to our task. For a fair comparison with previous work, as well as the tradeoff between model complexity and performance, we select the partially fine-tuned ResNet-50 as the CNN backbone to construct our DCVQE model.

Table 1. Performance comparisons of two series of architectures based on ResNet and ViT with full and partial fine-tuning strategies on KoNViD-1K dataset. Here ViT-B16/32 represents ViT base model with 16*16/32*32 input patch size, ViT-L32 represents ViT large model with 32*32 input patch size.

Models	SRCC	PLCC	KRCC	RMSE
ResNet-18$_{full}$	0.8893	0.8798	0.6955	0.2466
ResNet-18$_{partial}$	0.8888	0.8813	0.6979	0.2494
ResNet-50$_{full}$	0.8507	0.8407	0.6457	0.2889
ResNet-50$_{partial}$	0.9058	0.8933	0.7168	0.2308
ResNet-101$_{full}$	0.8511	0.8317	0.6365	0.2847
ResNet-101$_{partial}$	0.9075	0.8962	0.7166	0.2278
ViT-B16$_{full}$	0.8620	0.8818	0.6759	0.3570
ViT-B16$_{partial}$	0.7786	0.8103	0.5849	0.3587
ViT-B32$_{full}$	0.7716	0.8066	0.5788	0.3405
ViT-B32$_{partial}$	0.7639	0.8038	0.5707	0.3878
ViT-L32$_{full}$	0.7881	0.8131	0.5905	0.3267
ViT-L32$_{partial}$	0.8406	0.8708	0.6516	0.2786

Table 2. Performance comparisons of two models with ResNet-50 and PHIQNet feature extraction backbones. The tests are conducted on KoNViD-1K dataset.

Models	SRCC	PLCC	RMSE
ResNet-50$_{partial}$ + DCVQE	0.8382	0.8375	0.3515
PHIQNet + DCVQE	0.8376	0.8313	0.3599

2 Optimal Number of DCTr Layers

We conduct the ablation study on KoNViD-1K dataset to find out the optimal number of DCTr layers to construct our DCVQE model. As listed in Table 3, only one DCTr layer does not adequately solve the VQA problem, while stacking 3 DCTr layers significantly increases the performance. Further increases in layers do not improve performance. As a result, we set 3 as the optimal number of DCTr layers for our model.

Table 3. Performance comparisons of the different numbers of DCTr layers.

Layer #	SRCC	PLCC	RMSE
1	0.7954	0.8013	0.3688
3	0.8382	0.8375	0.3515
5	0.8346	0.8305	0.3592
7	0.8350	0.8332	0.3527

3 More Studies on the Proposed Correlation Loss

To find out how the proposed correlation loss additionally helps to improve NR-VQA, we apply the proposed losses to train the baseline Transformer and DCVQE models, respectively. The well-known pairwise ranking loss (PW-RL) is also involved in our study. Learned from subsection 4.4 of the paper, the best performance can be achieved with a temporal range selected from 9 to 15, so we only conduct the experiments under 3 different range settings of 9, 12, and 15. The test results are shown in Fig. 1, where we can see that no matter which architecture and temporal range are selected, the introduction of our correlation loss can consistently help to improve VQA performance. The PW-RL is also comparably well to optimize our DCVQE model. However, its solution reaches the highest RMSE. The reason is that the PW-RL will be converted to cross-entropy loss for training so that the optimization strength might be too strong for pairs with wrong ranking orders but small Mean Opinion Score (MOS)

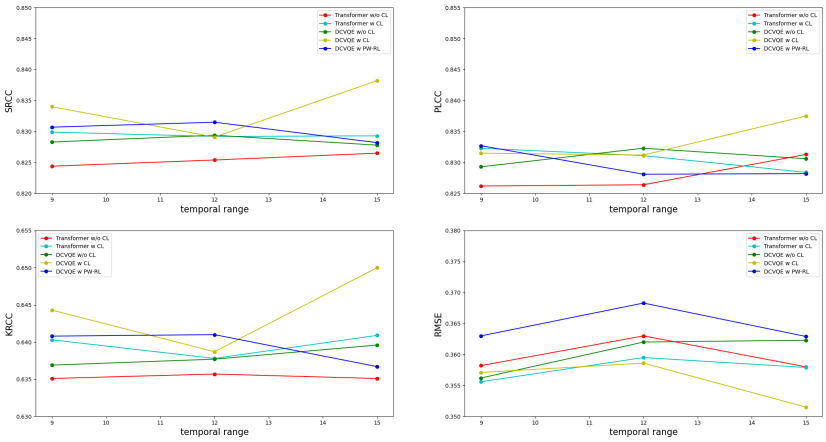

Fig. 1. Performance comparisons among 'with correlation loss' (w CL), 'without correlation loss' (w/o CL), and 'with pairwise ranking loss' (w PW-RL) under different models and settings on KoNViD-1K dataset.

differences. Fortunately, our correlation loss can better handle this situation because both ranking orders and MOS differences are considered.

Additionally, to show how the proposed correlation loss and architecture benefit real VQA tasks, we provide MOS prediction results of 4 KoNViD-1k sample videos in Table 4. From this table, we can see that (1) DCVQE$_{cl}$ maintains the order relation among the samples but DCVQE$_{l1}$ and Vanilla-Transformer fail, and (2) both the Mean Absolute Errors (MAEs) of DCVQE$_{cl}$ and DCVQE$_{l1}$ are lower than that of Vanilla-Transformer thanks to the new hierarchical architecture of DCVQE.

Table 4. MOS prediction results of 4 KoNViD-1k samples: DCVQE$_{cl}$ is trained with proposed loss (Eq. 5 of the paper); DCVQE$_{l1}$ is trained with L1 loss.

Video Id	Ground Truth	DCVQE$_{cl}$	DCVQE$_{l1}$	Vanilla-Transformer
5319047612	1.35	1.95	2.00	1.99
4265470174	1.56	1.96	1.96	1.97
3521396571	3.54	3.56	3.58	3.72
12893008605	3.55	3.58	3.55	3.68
MAE	-	0.26	0.27	0.34

4 Computational Cost Analysis

Compared with the Vanilla-Transformer, our DCVQE model has a lower computational cost. For example, to calculate the attention weights for one single frame, the time complexity of the Vanilla-Transformer is $O(DN)$, while that of our DCVQE is $O(D * \frac{N}{C})$ because an input video will be split into a number of clips for processing (where D denotes the dimension of feature, N is the total frame size and C is the clip number).