

Interpretable Long-term Action Quality Assessment - Supplementary File

BMVC 2024 Submission # 517

A Details of Implementations

Our temporal decoder module adopts a transformer decoder structure, utilizing queries with positional encoding to process sequences of video clips. The initialization of the query is adjusted by using `pytorch init.normal_` to change the variance, while keeping the mean unchanged. This decoder is equipped with 4 attention heads and 2 layers, enhancing its capability to handle temporal sequence data. Additionally, to maintain the integrity of information flow and to enhance feature integration, the decoder incorporates skip connections, which help prevent gradient vanishing issues in deep network layers, thereby improving the model's efficiency and accuracy in processing video data. Moreover, our decoder applies the ReLU activation function after each linear transformation and employs a dropout rate of 0.7 to enhance the model's generalization ability and reduce overfitting, ensuring stability and reliability when handling diverse video data.

The weight-score regression module is designed with a dual-pathway architecture for processing input through separate mean and weight calculations. It comprises linear transformations where the input x is first mapped from 1024 to 512 dimensions by `layer1_mean` and `layer1_weight`, and subsequently to 256 dimensions by `layer2_mean` and `layer2_weight`. The final layer, `layer3_mean`, computes a single value while `layer3_weight` uses a softmax function for weight computation, ensuring outputs are probability distributions. The class also uses ReLU activation functions for non-linear transformations. During the forward pass, the network computes weighted sums of features by multiplying mean values with corresponding weights from both pathways, integrating these to produce final logits.

B Additional Result

Effect of different variance to initialize query Table 1, 2, and 3 respectively show the SRCC results for using different variances to initialize query embeddings in RG, Fis-V, and Logo. The results demonstrate that larger variances lead to better SRCC results.

Effect of different modules Table 4 shows the results of our ablation studies on four subclasses of RG, demonstrating the effectiveness of each module. Among the results, it is observed that for all subclasses except the ball label, each module enhances the overall Spearman's Rank Correlation Coefficient (SRCC) results. It is particularly noteworthy that attention loss facilitates the most substantial improvement in SRCC results.

Variance	Ball	Clubs	Hoop	Ribbon	Avg.
0.1	0.8099	0.7963	0.7794	0.8368	0.8056
0.5	0.8144	0.8084	0.7628	0.8297	0.8038
1	0.8416	0.7839	0.7735	0.8413	0.8101
2	0.7978	0.8319	0.7736	0.8488	0.8130
3	0.8107	0.8272	0.7664	0.849	0.8133
5	0.7915	0.8042	0.786	0.8412	0.8057
10	0.8075	0.7782	0.8248	0.8128	0.8058

Table 1: Different variance for query initialization of RG dataset.

Variance	PCS	TES	Avg.
0.1	0.824	0.7008	0.7624
0.5	0.8199	0.706	0.7630
1	0.7906	0.674	0.7323
2	0.858	0.6954	0.7767
3	0.8408	0.7138	0.7773
5	0.6685	0.7165	0.6925
10	0.6667	0.6905	0.6786

Table 2: Different variance for query initialization of Fis-V dataset.

Variance	SRCC
0.1	0.7162
0.5	0.7173
1	0.7513
2	0.7156
3	0.6731
5	0.6766
10	0.6603

Table 3: Different variance for query initialization of LOGO dataset.

C Visualization of Attention Map

Variance in Query Initialization Module We compare different variances in the query initialization module as shown in Figure 1. The results show that using relatively larger variances can enhance network performance and increase the correlation in the self-attention map.

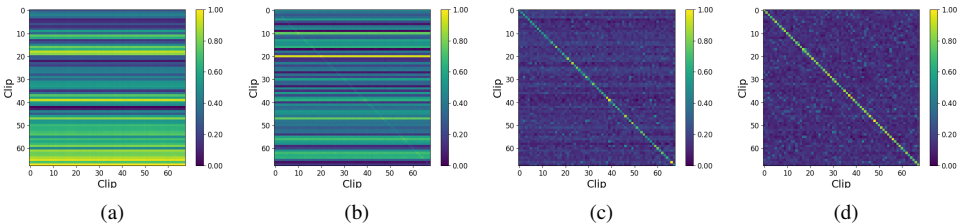


Figure 1: Self-attention map of query initialized with different variances. Figures 1(a), 1(b), 1(c), and 1(d) illustrate the self-attention maps corresponding to variances of 0.1, 1, 2, and 8, respectively.

Ball				
Variance	×	×	×	✓
Positinal Encoding	×	×	✓	✓
Attention Loss	×	✓	✓	✓
Results	0.4828	0.8201	0.8233	0.8233
Clubs				
Variance	×	×	×	✓
Positinal Encoding	×	×	✓	✓
Attention Loss	×	✓	✓	✓
Results	0.6346	0.813	0.8089	0.8524
Hoop				
Variance	×	×	×	✓
Positinal Encoding	×	×	✓	✓
Attention Loss	×	✓	✓	✓
Results	0.6264	0.7606	0.7696	0.837
Ribbon				
Variance	×	×	×	✓
Positinal Encoding	×	×	✓	✓
Attention Loss	×	✓	✓	✓
Results	0.7669	0.8353	0.8366	0.8568
Avg.	0.6277	0.8073	0.8096	0.8424

Table 4: Ablation study on RG of four subclasses.