TEMPORAL ATTENTION NETWORKS FOR NO-REFERENCE VIDEO QUALITY ASSESSMENT

Joshua Peter Ebenezer

University of Texas at Austin

ABSTRACT

This project focuses on exploring the use of attention models to extend the use of no-reference (NR) image quality assessment (IQA) algorithms for the task of NR video quality assessment (VQA). Attention networks are trained on features extracted using state-of-the-art IQA algorithms. A deep learning architecture for VQA is proposed, using a pretrained network trained for IQA. Results are mixed, and show that while attention can improve performance on some tasks, current VQA databases may not be large enough for end-to-end training.

***Index Terms*—** Attention, video quality, deep learning

1. INTRODUCTION

IQA algorithms have had great success in predicting human subjective responses to image quality even without the presence of a reference image. The problem of NR VQA, on the other hand, is very much an open area of research. State-of-the-art (SOTA) VQA algorithms typically incorporate spatial and temporal features. Pooling these features across time and across frames can be done in a number of different ways. Recent work [1] suggests that the efficacy of pooling methods is highly content-dependent. Datasets which do not have much temporal variation in the frame quality, and for which the quality of each frame is more or less constant throughout the video, are more likely to be agnostic to the pooling method. In such cases, taking the average of frame-wise scores or features would give reasonable performance. On the other hand, humans are known to have a persistent memory of sudden dips in quality, and give higher importance to the quality of frames towards the end of the video and at the beginning. Several models have been suggested for this, including the temporal hysteresis [2] model and models of primacy and recency effects. Attention networks [3] offer a flexible way of modeling these effects in a learnable manner. These networks can learn to weigh frames differently based on their contribution to the final score, and can be seamlessly integrated into deep networks. On datasets that do not have much variation in frame quality over time, attention networks can still dynamically learn to give equal attention to every frame, while in other cases attention networks may be able to model human responses given enough data.

When framewise features are used to estimate video quality, the most common approach has been to use the average of the features across frames to train a learning engine that maps the average feature vector to a mean opinion score collated from a subjective human study. The success of deep learning in a number of tasks, including image quality assessment, now brings up the question of how these deep networks can be used for the task of video quality assessment, in ways other than averaging framewise features while still leveraging their success on image quality assessment tasks. A possible solution could be attention networks, which can be seamlessly integrated into pretrained deep networks for image quality assessment to equip them to perform the task of video quality assessment.

Features were extracted from SOTA IQA algorithms and fed to an attention network that tries to learn which frames to pay attention to while determining the final video quality score. The learning engine used on the output of the attention network is a 3-layer neural network. Results were compared against using the average of the frame-wise features to train the same learning engine. Note that since this is a study of the effectiveness of attention as a pooling strategy, and not a study on comparing learning engines such as neural networks and SVRs, the same learning engine was used for training both the averaged feature vectors and the output of the attention network. A end-to-end deep network is also proposed, making use of a state-of-the-art deep learning network for image quality assessment.

2. RELATED WORK

2.1. IQA algorithms

Many IQA algorithms have achieved impressive performance on IQA databases. Some of these are perceptually driven and based on statistical models of natural images, while others try to use low-level features to quantify distortions. BRISQUE [4], NIQE [5], HIGRADE [6], FRIQUEE [7], and CORNIA [8] are SOTA NR-IQA algorithms, and have been chosen for this project to act as feature extractors for an attention network. NIQE outputs only a single score, but in this project the features used in NIQE are also extracted for training. Paq-

2-Piq [9] is a deep network trained on a massive database of image and patch quality scores, and has been shown to outperform classical IQA algorithms on large databases. A pre-trained Paq-2-Piq is used with attention to build a deep network for the task of VQA.

2.2. Dataset

The Konvid-1k database [10] is one of the largest databases for VQA that has been publicly released and was used for this project. Flickr-Vid 150k [11] is a much larger database, specifically created for use in developing deep networks for VQA, but has not been publicly released yet.

2.3. VQA algorithms

VQA algorithms such as MOVIE [12] and Video-BLIINDS [13] often use spatial features or frame-wise temporal features, such as features based on frame-differences. These features are typically averaged across frames. Attention provides an intuitive and data-driven way to weigh these features differently, and may give rise to better-performing modifications of these VQA algorithms if given enough data.

3. METHODOLOGY

3.1. Extension of classical IQA

A self-attention model was used, based off work by Bahdahnau et al. [14] on the task of machine translation. Given a feature vector for each video $\mathbf{x}$ of dimension $N \times M$, where M is the dimension of the feature vector for each frame computed by an IQA algorithm and N is the number of frames, the attention model computes key ($\mathbf{k}$), query($\mathbf{q}$), and value($\mathbf{v}$) matrices for each video.

$$\mathbf{k} = \tanh(\mathbf{x}\mathbf{W_k}) \tag{1}$$

$$\mathbf{q} = \tanh(\mathbf{x}\mathbf{W_q}) \tag{2}$$

$$\mathbf{v} = \tanh(\mathbf{x}\mathbf{W_v}) \tag{3}$$

where $\mathbf{W_k}$, $\mathbf{W_q}$, $\mathbf{W_v}$ are trainable weight matrices of dimension $M \times A$. Hence the key, query, and value matrices have size $N \times A$. These are used to generate feature-dependent weights which are the dot products between the rows of q and k.

$$\mathbf{e} = \mathrm{diag}(\mathbf{q}\mathbf{k}^\mathbf{T}) \tag{4}$$

$\mathbf{e}$ has dimensions $N \times 1$, and is normalized with softmax.

$$\mathbf{s}_i = \frac{\exp(\mathbf{e}_i)}{\Sigma_i \exp(\mathbf{e}_i)}\ , i = 1, 2...N \tag{5}$$

The output of the attention layer is

$$\mathbf{o} = \mathbf{v}\mathbf{s}^T \tag{6}$$

which is of size $A \times 1$. The output represents a feature vector of size A that is a weighted combination of each frame's feature vectors, and has been weighted by the network depending on its contribution to the final score. 4 such attention heads are used, and the outputs concatenated to form a vector of length $4A$ that represents each video. This is then fed to the learning engine.

3.2. End-to-end deep learning

A pre-trained Paq-2-Piq RoiPoolModel was used as a base for a video quality assessment network. It is computationally prohibitive to train all frames in a video at once, and research has shown that correlation between adjacent frames is so high that it may not be necessary to do so [15]. N equally spaced frames are selected from each video. Each frame is passed through the body of a pre-trained Paq-2-Piq network. The output of the body of Paq-2-Piq is a 1024 length feature vector for each frame. The $N \times 1024$ vectors are fed to a uni-directional LSTM, which generates an abstract description of all the frames of length $N \times 1024$. The output of the LSTM is fed to an attention network, which in turn feeds 3 fully connected layers. The attention network used here is a self-attention multihead model, introduced by Vaswani et al. [3]. The entire model was finetuned on Konvid-1k.

4. EXPERIMENTS

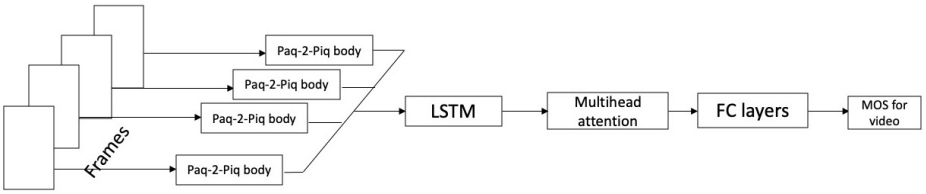
4.1. Extension of classical IQA

The experimental setup for the extension of classical IQA is different from that used for end-to-end learning. This section deals with the former.

4.1.1. Dataset

The dataset used was Konvid-1k [10]. 80% of the data was used for training, 10% for validation and 10% for testing. BRISQUE and NIQE are relatively fast and do not occupy much space, hence they were computed on all frames of all videos in Konvid. CORNIA is fast but the size of the feature set is large (20,000 features for each image) while FRIQUEE is very time consuming. It has been shown [15] that it is not necessary to extract features from every frame in a video for reasonable performance, and that performance from taking features from only 1 frame every second is similar to the performance if more frames are used to extract features from. Hence for FRIQUEE, HIGRADE, and CORNIA, only features from every 20 frames were extracted, while for BRISQUE and NIQE all frames were used. This ensured that at least one frame from each second of each video is represented, because the minimum frame rate of the videos in Konvid is 24 fps.

Fig. 1. Architecture for end-to-end training



4.1.2. Variable length problem

Each video has a different number of frames, and this poses a challenge because neural networks expect fixed-size inputs in each batch. Hence for each video, 4 groups of contiguous frames are randomly selected such that they are all of the same length, equal to half of the length of the shortest video. All groups are assigned the same overall mean opinion score (MOS) for the video. In this way, the amount of data increases four-fold, and the problem of variable lengths is resolved, but this is under the assumption that these groups of frames are large enough to represent the MOS of the whole video individually. During testing time, 4 groups of frames are selected and the MOS predicted for each of them are averaged to form the final predicted MOS.

4.1.3. Hyperparameter search

The network was trained for a maximum 300 epochs with a hyperbolic-decaying learning rate, with an early stopping patience of 40 epochs. In other words, if the validation loss fell for 40 epochs, the model terminated training and restored its parameters from the point at which the validation loss started to fall. This is a heuristic found from observing the training and validation loss curves. The Adam optimizer was used. The learning rate was a hyperparameter found through grid search (described later). The learning engine is a three layer neural network, with the ReLU activation function and dropout at each layer.

The architecture can be described as follows. If the dropout at the input layer is d, at the second layer it is $2d$ and at the output layer it is $3d$. If the number of nodes at the input layer are n, at the second layer they are $n/2$ and the third $n/4$. This offers a flexible way of searching for hyperparameters, and it is common practice to have dropout increasing in deeper layers and number of nodes decreasing, in order to force the network to learn a compact representation before it regresses on the MOS. The dropout nearer the input layers is always lower, in order that the network is able to retain enough information at the source to generate a meaningful output. Using this notation, a grid search is performed on the following candidates for learning rate, dropout, number of nodes, and attention nodes A (defined earlier).

These sets of values were chosen heuristically, after a lot of trial-and-error found that reasonable results could be found in this range. Note that the attention nodes are a part of the grid search only when the attention network was being tested. When framewise feature vectors are simply averaged, attention is not used and A is not a part of the grid search, but the other hyperparameters are. The best set of hyperparameters for each algorithm was chosen on the basis of best performance on the validation set.

Table 1. Hyperparameter grid search candidates

Hyperparameter	**Candidates**
DROPOUT d	0.1,0.01,0.001,0.0001
NUMBER OF NODES n	64, 128,256
LEARNING RATE	0.1,0.01,0.001,0.0001
ATTENTION NODES A	8,16,32

4.2. End-to-end learning

A pre-trained Paq-2-Piq RoiPoolModel was used as a base for a video quality assessment network. 20 equally spaced frames were selected from each video in Konvid. Frames were processed exactly as they were for Paq-2-Piq training, i.e. rescaling by division by 255 and conversion to float were the only operations performed. Each frame was passed through the body of a pre-trained Paq-2-Piq network. A number of different architectures were tested for the part of the network following this. The best performing architecture is show in Fig. 1.

About 200 different architectures and hyperparameters were tested for end-to-end training, but most configurations were found to be highly unstable. The network overfits very easily. Reducing the capacity by introducing regularization, dropout and removing layers does not solve the problem, and the network either overfits or underfits badly, without being able to generalize well. This is typical when the dataset is too small relative to the difficulty of the problem. End-to-end training without attention was found to be very unstable as well. Best results were obtained when Paq-2-Piq was trained with a learning rate of 1e-5 and the layers above it trained with a rate of 1e-4, with the Adam optimizer. The learning rate was set to exponentially decay in all experiments.

In another experiment, frame differences were computed

and fed to a parallel network which was pre-trained on ImageNet. The features were concatenated with the features obtained by Paq-2-Piq, and fed to the LSTM and attention networks. This did not give significantly improved results on the validation set.

Experiments were also conducted with freezing the Paq-2-Piq pre-trained model and only training the attention and FC layers. This did not do better than end-to-end finetuning.

5. RESULTS

5.1. Extension of classical IQA

Table 2 has test SROCC values for various IQA algorithms when the frames are pooled with attention and with averaging. Attention performs better than frame-wise averaging for BRISQUE, NIQE and FRIQUEE, and achieves comparable performance for CORNIA and HIGRADE.

Visualization and interpretation are very useful properties of attention. Fig. 2 shows four plots of attention weights over time and one plot of NIQE scores for that segment. These weights were obtained from a model trained on NIQE features and the NIQE score. NIQE scores quantify deviations from the natural scene models, and so a higher NIQE score indicates lower quality. Notice that the weights go up around any area where there is a drop in quality. The drop in quality between frames 20 and 40 is almost mimicked by the first two sets of attention weights, which give those frames a higher weightage. At frame 60 there is an increase in the quality (shown by a sharp dip in the NIQE score) and this dip is represented by attention weight vector 4. It is clear from this that the attention network has (in this case) learnt to pay more attention to poor quality frames and pay less attention to high quality frames. Each of the different sets of attention weights can learn to focus on different temporal variations.

5.2. End-to-end training

The best architecture had a SROCC of 0.61 on validation data and 0.47 on test data. These are not the state-of-the-art and clearly show the necessity of creating larger VQA datasets for end-to-end training of deep networks. Training and validation curves are plotted in Fig. 3.

6. DISCUSSION

Attention networks give mixed results on the Konvid-1k database. For most of the classical IQA algorithms, attention is able to show a marked increase in performance over frame-wise averaging. Visualization and interpretation of the pooling from attention is a very useful tool to get into the black boxes of deep networks. On the other hand, end-to-end training is unstable, both with and without attention, even with the use of a pretrained SOTA model like Paq-2-Piq.

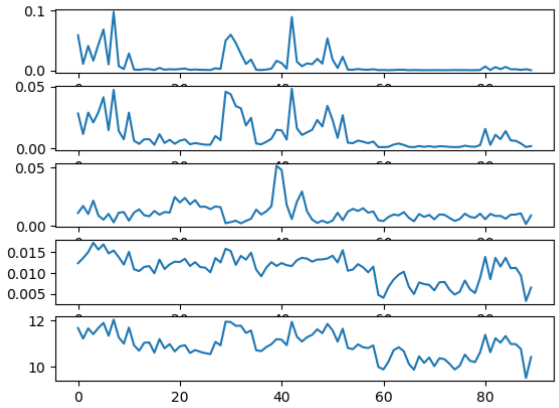
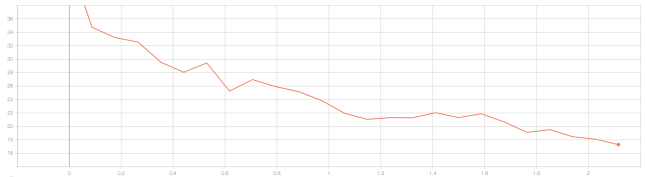
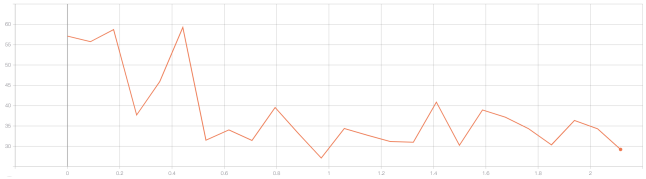


Fig. 2. Attention weights and NIQE scores over time. The first four plots are 4 different weighting vectors **s**, and the last shows the NIQE scores for each frame over time.



(a) Training loss over time



(a) Validation loss over time

Fig. 3. Loss figures.

Hyperparameter search is intensive and many configurations do not converge well. This typically happens when datasets are not large enough relative to the difficulty of the problem. These results tentatively point to the efficacy of attention as a temporal pooling strategy, but larger databases are required to provide better validation of this hypothesis. Another issue is that Konvid-1k has videos whose quality remain relatively constant over time. Temporal variation in quality is minimal, and attention may do no better than averaging if there is no variation at all. Nevertheless, attention networks seem to have captured even these minor variations, as can be seen in their improvement over frame-wise averaging. Attention networks may be even more powerful when trained and evaluated on large datasets of videos with stalling events, where temporal variations in quality are clearly perceived.

Table 2. SROCC values on Konvid-1k database with a 3 layer neural network

IQA algorithm	Frame-wise average	With attention
BRISQUE	0.2978/0.3717	**0.5993/0.6139**
NIQE (features+score)	0.5908/0.5978	**0.6609/0.6685**
FRIQUEE	0.5916/0.5993	**0.6549/0.6451**
HIGRADE	**0.7258/7267**	0.7011/0.6671
CORNIA	**0.7485/0.7435**	0.7148/0.7152

7. CONCLUSION

This project has studied the use of temporal attention networks for the extension of NR IQA algorithms to the task of NR VQA. Attention has been shown to give a slight increase in the performance of most SOTA IQA algorithms when extended to VQA. However, when attention is integrated into a deep network for end-to-end training, it is found the largest existing VQA database may not be large enough for the task, and that training is unstable. There is thus a need to develop and release larger databases for VQA for end-to-end training of deep networks for VQA. The success of attention on classical algorithms is promising, and attention could be used in any existing VQA algorithm as a temporal pooling strategy. A strong argument for attention can be especially made for the problem of VQA for videos distorted by stalling or buffering events, and this could be an avenue for future work when large databases for buffering distortions are made available.

8. REFERENCES

[1] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," *arXiv preprint arXiv:2002.10651*, 2020.

[2] Kalpana Seshadrinathan and Alan C Bovik, "Temporal hysteresis model of time varying subjective video quality," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 1153–1156.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[5] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[6] Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans, "No-reference quality assessment of tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.

[7] Deepti Ghadiyaram and Alan C Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of vision*, vol. 17, no. 1, pp. 32–32, 2017.

[8] Peng Ye and David Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 2012.

[9] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," *arXiv preprint arXiv:1912.10088*, 2019.

[10] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[11] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe, "No-reference video quality assessment using multi-level spatially pooled features," *arXiv preprint arXiv:1912.07966*, 2019.

[12] Kalpana Seshadrinathan and Alan Conrad Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[13] Michele A Saad, Alan C Bovik, and Christophe Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] Jari Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.