CVlab Progress

"What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment"

- Parmar and Morris

- Tries to work with the problem of action quality assessment
- AQA aims at assigning a score to a performed task
- Is primarily used to imitate human judges in athletics
- Can be used to judge skill level of painting / surgery / progress in physical rehabilitation
- Tries to quantify the quality of action

The authors state that most prior work done in this field involves training a network to learn features that serve only one task - estimating the final score

The authors pose the following question :
"can learning to describe and commentate on the action instances help improve the performance on the AQA task?"

More specifically, the authors speculate that training the network to do well in commentating on the action performed and recognizing the action in addition to score the quality of action helps it to produce better scores

Formally , the approach of the authors is:

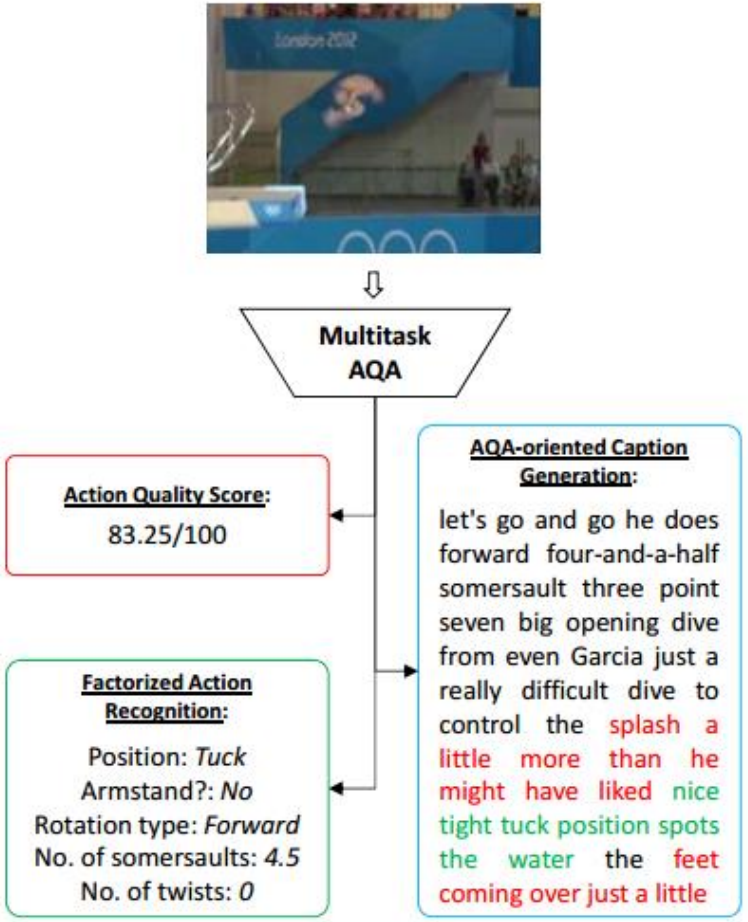
1.Utilizing 3D CNN's to learn spatio-temporal representation of salient motion and appearance

2.Use these representations to train a network / optimizer a loss function accounting for -

a) The action quality score
b) Factorized action classification
c) Generate verbal commentary of performance

All this will be done in an end to end manner

Optimizing the network to do well in these related tasks forces it to learn more generalized features in the lower layers and perform better



Multi-Task Learning

- MTL is a machine learning paradigm in which a single model caters to more than a single task.
- MTL tasks are generally chosen such that they are related to one another
- Networks have a common body that branches into task-specific heads
- The total network loss is the sum of individual task losses.

The network is able to learn richer representation in the common body section since it must be able to serve/explain all tasks.

MTL has been shown to gain improved generalization as compared STL

Multi Task learning in AQA

- Main task is to assess the action quality (AQA score)
- Auxiliary tasks are :
  - Action recognition
  - Caption / commentary generation
- Action recognition in turn consists of 5 fine-grained sub tasks
  - Recognizing position
  - Recognizing rotation type
  - Detecting armstand
  - Counting somersaults
  - Courting twists
- The network has a common body and 3 task dedicated heads.

Multi Task learning in AQA

AQA is modelled as a regression problem. Thus, the loss function for AQA :

$$\mathcal{L}_{AQA} = -\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 + |x_i - y_i|$$

where x_i is the predicted score and y_i is the ground truth score for each of the N samples.

For action recognition, the loss function is defined as :

$$\mathcal{L}_{Cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{sa}\sum_{j=1}^{k_{sa}} y_{i,j}^{sa} log(x_{i,j}^{sa})$$

Negative log likelihood is used as the loss function for the captioning task : $\mathcal{L}_{Cap} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{sl} ln(x^{cap}{}_{y^{cap}})$

The overall objective function to be minimized is the summation of all the losses

$$\mathcal{L}_{MTL} = \alpha\mathcal{L}_{AQA} + \beta\mathcal{L}_{AR} + \gamma\mathcal{L}_{Cap}.$$

Architecture

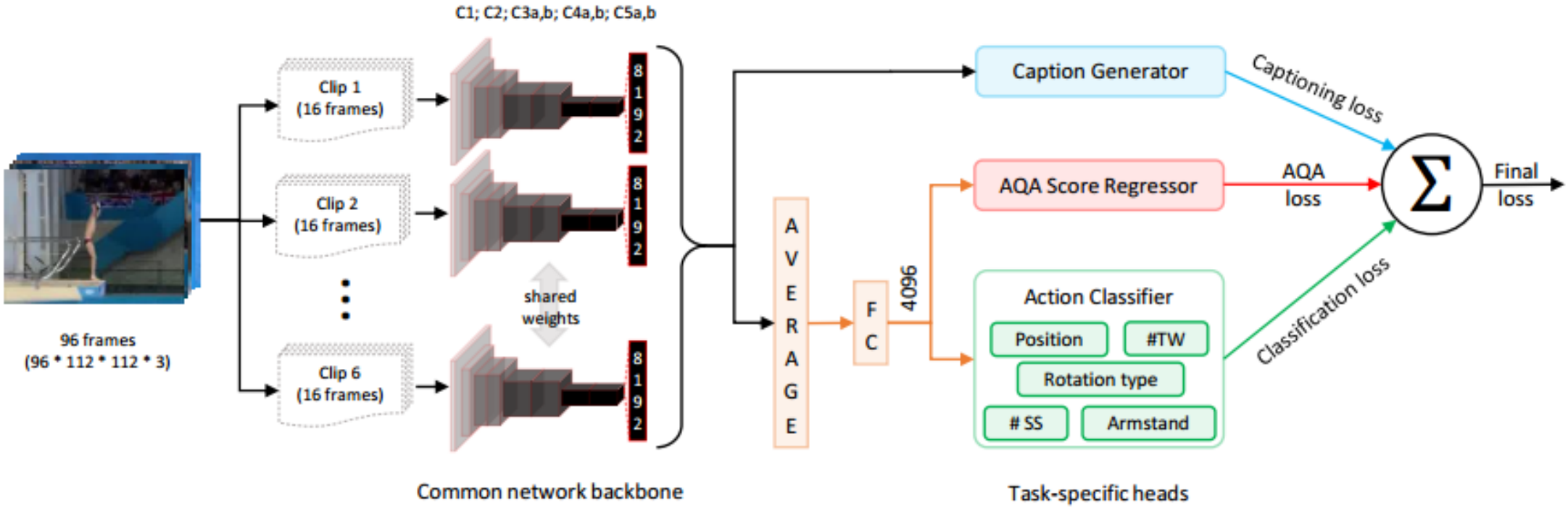
A 3D CNN is used to learn spatio-temporal representations from input video

- 96 frame video requires excessive memory to be analyzed by 3D CNN
- Video divided in small clips of 16 frames
- Representation extracted for each clip using 3D CNN
- Representation aggregated and used for generating video level representation

Two segments:

1. Common network backbone
2. Task specific head

Backbone learns shared representations and these representations are further processed through task specific heads to obtain more task-oriented features and outputs.



Averaging as aggregation (C3D-AVG)

Averaging as aggregation (C3D-AVG)

- Backbone consists of C3D network upto the fifth pooling layer.
- Average is used as a mean for aggregation. This might be an area for improvement
- Up to Average layer can be considered as an encoder, which encodes input video-clips into representations
- When averaged , these representations would correspond to the total AQA points gathered by the athlete
- subsequent layers can be thought of decoders for individual tasks.
- Since captioning is a sequence-to-sequence task, the individual clip-level features are input to the captioning branch before averaging

Multiscale Context Aggregation with Dilated Convolutions (MSCADC)

- Another architecture for the same task.
- 96 frame video is downsampled in one 16 frame clip
- Network backbone : C3D network is used, last 2 pooling layers are replaced with dilated convolution with a dilation rate of 2 . Batch normalization is introduced
- Task-specific heads: Heads consist of a context net followed by a few additional layers. The context net is where the feature maps are aggregated at multiple scale

	(Common network body)	
	C3(32); BN MP(1,2,2) C3(64); BN MP(2,2,2) {C3(128); BN} x2 MP(2,2,2) {C3(256); BN} x2 {C3(d=2,256); BN} x2 Dropout(0.5)	
	(Task-specific heads)	
(AQA Score Head)	***(Action recognition Head)***	***(Captioning Head)***
C1(12) {Cntxt net} MP(2,2,2) C3(12); BN	C1(12) {Cntxt net} MP(2,2,2) C3(12); BN	C1(12) {Cntxt net} MP(2,2,2) C3(12); BN
C3(1) AP(2,11,11)	***(Action recognition sub-heads)***	Enc. GRU Dec. GRU

Results

Tasks	C3D-AVG	MSCADC
AQA	89.60	84.72
+ Cls	89.62	85.76
+ Caps	88.78	85.47
+ Cls + Caps	**90.44**	**86.12**

Table 4: **STL vs. MTL across different architectures**. Cls - classifiction, Caps - captioning. First row shows STL results, while the remaining rows show MTL results.

# samples	1059	450	280	140
STL	89.60	77.27	69.63	64.17
MTL	**90.44**	**83.52**	**72.09**	**68.16**

Table 8: **STL vs. MTL generalization**. Training using increasingly reduced no. of training samples.

Method	Sp. Corr.
Pose+DCT [18]	26.82
C3D-SVR [16]	77.16
C3D-LSTM [16]	84.89
Ours MSCADC-STL	84.72
Ours C3D-AVG-STL	**89.60**
Ours MSCADC-MTL	86.12
Ours C3D-AVG-MTL	**90.44**
Segment-specific methods (train/test on UNLV Dive [16])	
S3D (best performing in [26])	86.00
Li *et al.* [13]	80.09
Ours MSCADC-STL	79.79
Ours C3D-AVG-STL	83.83
Ours MSCADC-MTL	80.60
Ours C3D-AVG-MTL	**88.08**

Table 7: **Performance comparison with the existing AQA approaches**.

Areas of Improvement

- The authors used average as a mean of feature aggregation in C3D-AVG-MTL network. We argue there might be better approaches, for example using a fully connected layer or some other trainable layer, which can improve the performance of the network
- The backbone network based on C3D used pretrained weights trained on UCF-101 dataset. A better approach might be using a C3D network with weights pretrained on kinetics dataset as the kinetics dataset is much larger. It has been shown that 3D convents trained on kindetics provide better results than other available video datasets.

Thank you