Attention based caption augmented W2VV++ Adhoc Video Search (AVS) trecvid task

Rahul Sharma[1,2], Deepak Mishra[2], Haresh Bhatt[3]
[1] rahul@sac.isro.gov.in, DECU, ISRO Ahmedabad, India
[2]deepak.mishra@iist.ac.in, Avionics Department,IIST, Thiruvananthapuram, Kerala, India
[3]haresh@sac.isro.gov.in, SAC, ISRO Ahmedabad, India

Abstract

In this paper we summarize our TRECVID 2020 video retrieval. We participated in Ad-hoc Video Search (AVS) task. For the AVS task, we developed our solutions based on W2VV++, a super version of Word2VisualVec (W2VV) by attempting optimization of hyperparameters and further augmenting it with attention based caption generation based text to text matching.

1. Approach

An attempt is done to augment the state of the art W2vv++ implementation. The w2vvpp model which won the 2018 Trecvid and set the change towards concept-less video. Firstly, experimental optimization of hyper parameters and different optimisers were tried and secondly, Query to captions similarity was explored to re-rank the outcome of the w2vv++.

1.1 Model optimization

Attempt is done to improve training performance of the W2vv++ model. Multiple optimisers were experimented and learning rate values and strategies used.

a. For various Optimizers

The W2vvpp model is trained using different optimizers. Following optimizer techniques are applied

- RMSprop
- Adam
- Weighted Adam
- Adagrad
- Adamax

There seems to be scope of further optimizations using Adamax and Adagrad as the model is further trainable.

b. For different learning rate (Strategies)

In the existing SOTA work of w2vvpp model, learning rate strategy is adapted as step wise reduction after 3 consecutive fall, and early stop after 10 such sequential events. Attempted few alternate learning rates and its reduction techniques as the model while trailing stops learning and approaches early stop by around 20th epoch. Only marginal improvements in MaP were seen at the cost of increasing the training/learning epochs.

1.2 Caption based w2vv++ augmentation

For Show and Tell implementation, MSCOCO 14 dataset and as encoder a pretrained ResNet-101 model is used for training the attention-based caption generation. Where as, in W2VV++ the training is done on a joint collection of MSR-VTT and TGIF for video representation and deep visual features are extracted per frame by pre-trained CNN models ResNet-152 & ResNeXt-101.

Attention based captions are generated separately based upon pytorch implementation of Show and Tell paper [6]. Captions *(Cv)* are generated for keyframes sampled every 0.5 seconds from V3C1 collection. Sentence embeddings of these captions *a(ci)* are then obtained using W2VV++ multi scale embedding generation.

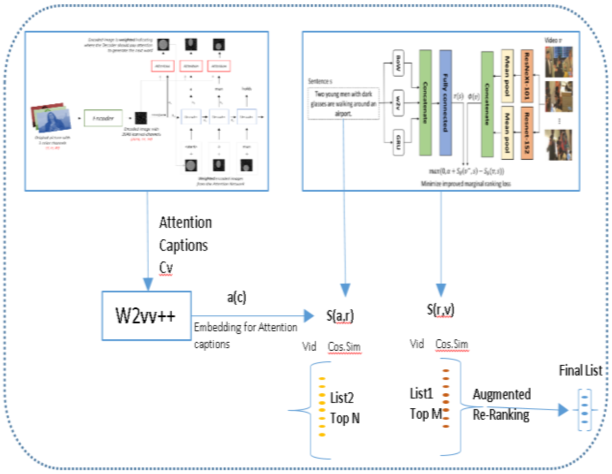

Figure 1: Conceptual diagrams of Augmented W2VV++ model.

Further, For each of *a(ci)*, cosine similarity between query sentence embedding and respective video frame caption, is calculated as S(ai,r). The list1 thus obtained of top N (10,000) sorted cosine similarities is used to augment W2VV++, list2 of top N (1000) sorted cross similarities S(r,v) between query sentence embedding r(s) and respective video embedding i.e. Ø[v_i]. Figure-1 details the entire approach conceptually. Resulting augmented Re-ranked list is thus submitted as the outcome Run of the experiment. In summary

a. Check vid in each of pair (vid, S(r,vi)) of list2 with that of vid in list1 pair (vid, S(ai,r)).
b. If vid matches, improve the corresponding S(r,vid) in list2 by 10%.
c. If vid doesn't match, retain the list2 values.

2. Our Results

The optimization of W2VV++ model has results as jn Table 1.0. The Adamax optimizer and varying learning rate seems promising for better mAP.

Table 1.0: Model Optimization

Epochs	Optimizer	mAP
19	RMSProp	56.10
16	Adam	55.73
17	AdamW	55.10
22	Adagrad	45.33
27	Adamax	56.23

As depicted in the Figure-2, in trecvid 2020 AVS results, the augmented W2VV++ submitted run scored an mAP of 0.107.

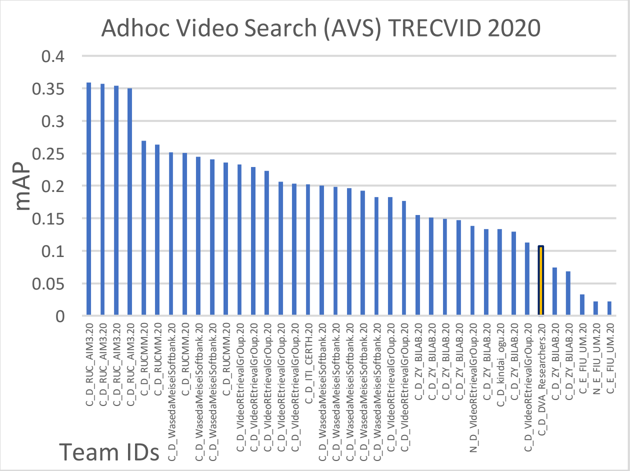

Figure 2: TRECVID 2020, AVS scores

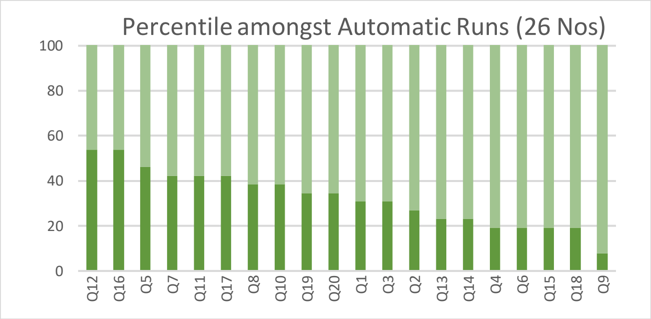

Figure 3: Query-wise Percentile amongst automatic runs (26 Nos)

Figure-3 details querywise performance of the total 26 automatic runs, as detailed in Trecvid 2020 overview paper [7].

Acknowledgments

The authors are grateful to NIST and the TRECVID [9] coordinators for the benchmark organization effort. We also acknowledge the support extended by DECU/ISRO and IIST/DOS for this work. For the availability of ResNext model we thank Li Xirong (Associate Professor) Renim University.

References

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[2] Niluthpol Chowdhury Mithun et. al. learning Joint Embedding with Multimodal ues for Cross-Modal Video-Text Retrieval, ICMR'18, June 11–14, 2018

[3] Xirong Li, et. al. W2VV++: Fully Deep Learning for Ad-hoc Video Search MM'19 ACM, October 21–25, 2019

[4] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In Proceedings of TRECVID 2018. NIST, USA, 2018.

[5] Kelvin Xu et. al. , Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

[6] Sagar Vinodababu GitHub repositories for Pytorch implementation of Show & Tell paper.

[7] George Awad and Asad A. Butt and Keith Curtis and Yooyoung Lee and Jonathan Fiscus and Afzal Godil and Andrew Delgado and Jesse Zhang and Eliot Godard and Lukas Diduch and Jeffrey Liu and Alan F. Smeaton and Yvette Graham and Gareth J. F. Jones and Wessel Kraaij and Georges Quénot, TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains, Proceedings of TRECVID 2020,2020,NIST, USA.

[8] Rossetto, Luca and Schuldt, Heiko and Awad, George and Butt, Asad A, V3C--A Research Video Collection,International Conference on Multimedia Modeling (349-360) ,2019, Springer

[9] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722