

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

Rahul Sharma^{1,2}, Deepak Mishra², Haresh Bhatt³

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-

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 cues 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.