Summarizing Videos with Attention

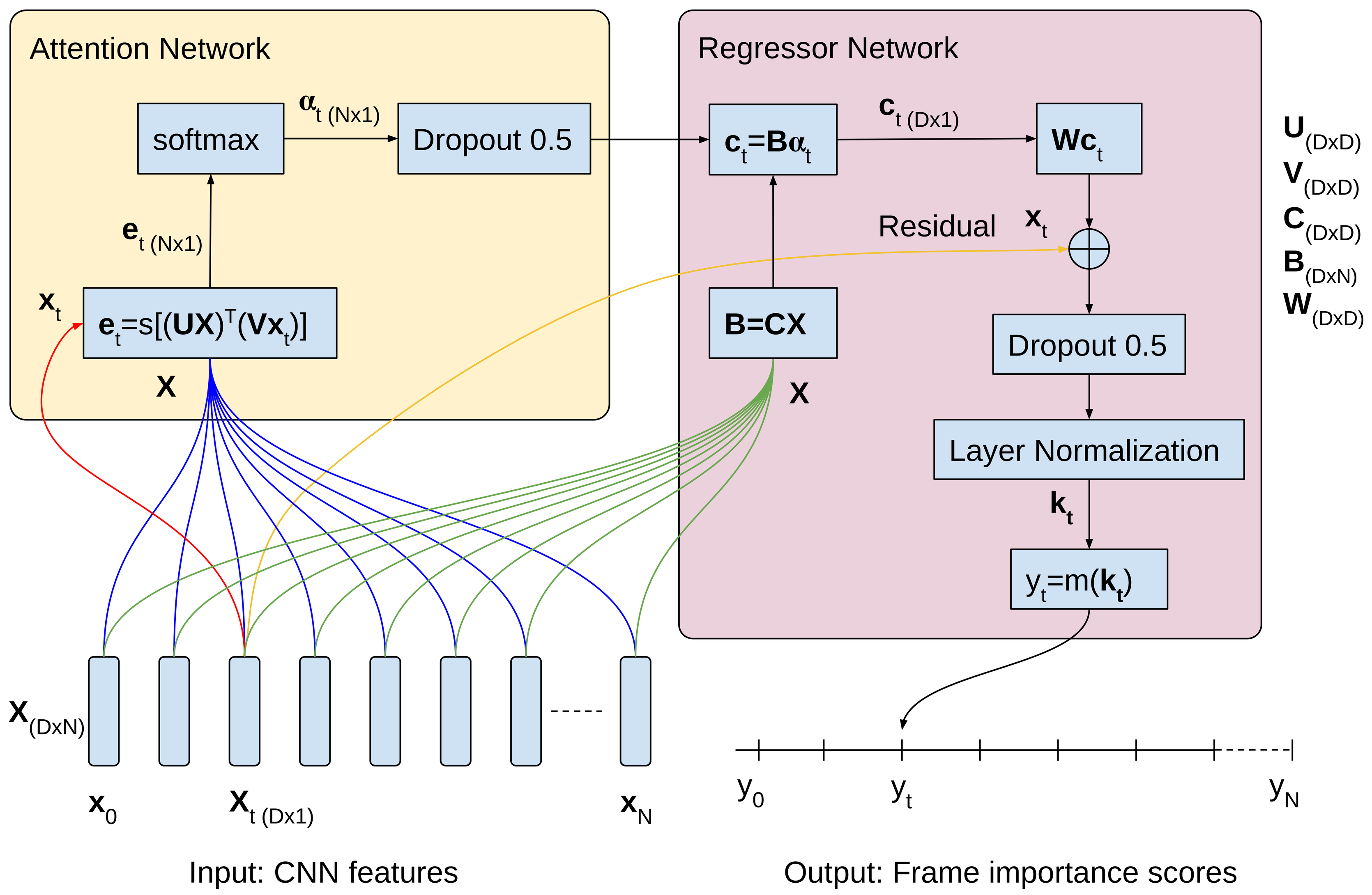
Jiri Fajtl[1], Hajar Sadeghi Sokeh[1], Vasileios Argyriou[1], Dorothy Monekosso[2] , Paolo Remagnino[1]
[1]Kingston University, London, UK [2]Leeds Beckett University, Leeds, UK
{J.Fajtl, h.sadeghisokeh, Vasileios.Argyriou, P.Remagnino}@kingston.ac.uk, D.N.Monekosso@leedsbeckett.ac.uk
Robot Vision Team (RoViT)

ACCV 2018 2 – 6 December 2018 Perth Western Australia
AIU2018 workshop

Introduction

- A new technique for supervised, keyshots video summarization
- Based on a novel soft, self-attention model for sequence to sequence transformation
- Shows superior performance compared to the current state of the art

Network Architecture

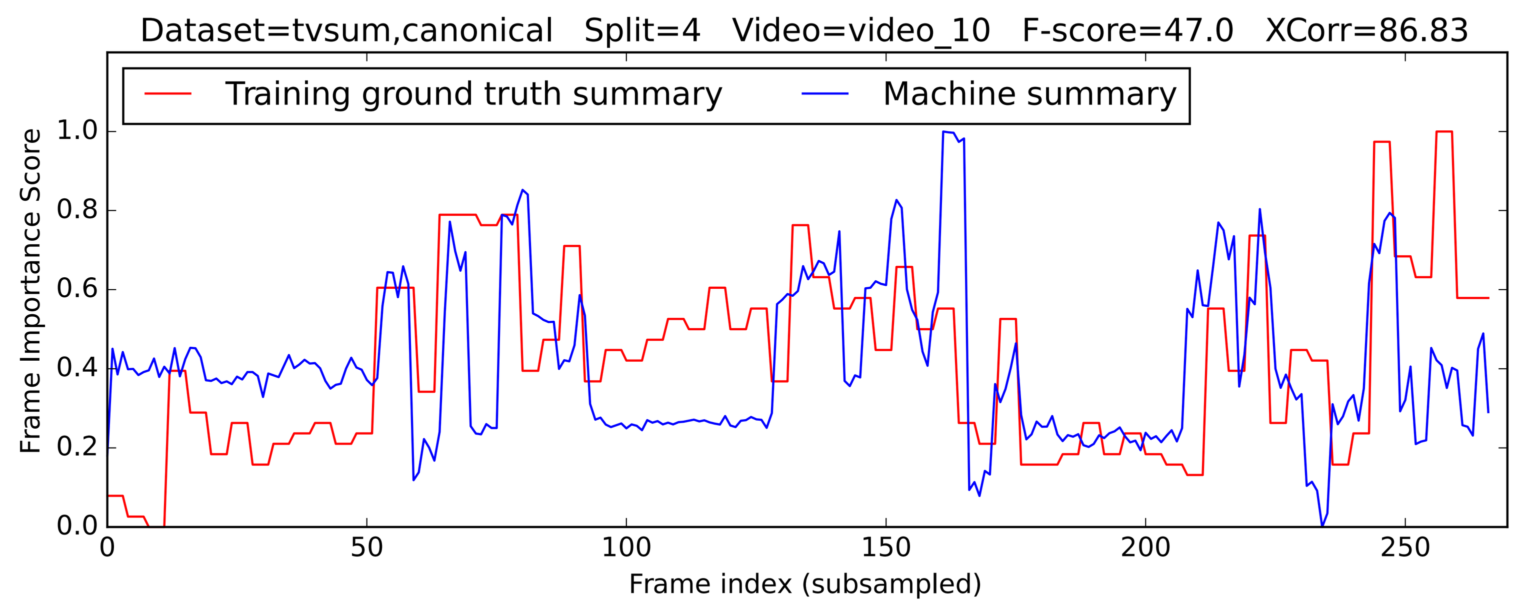


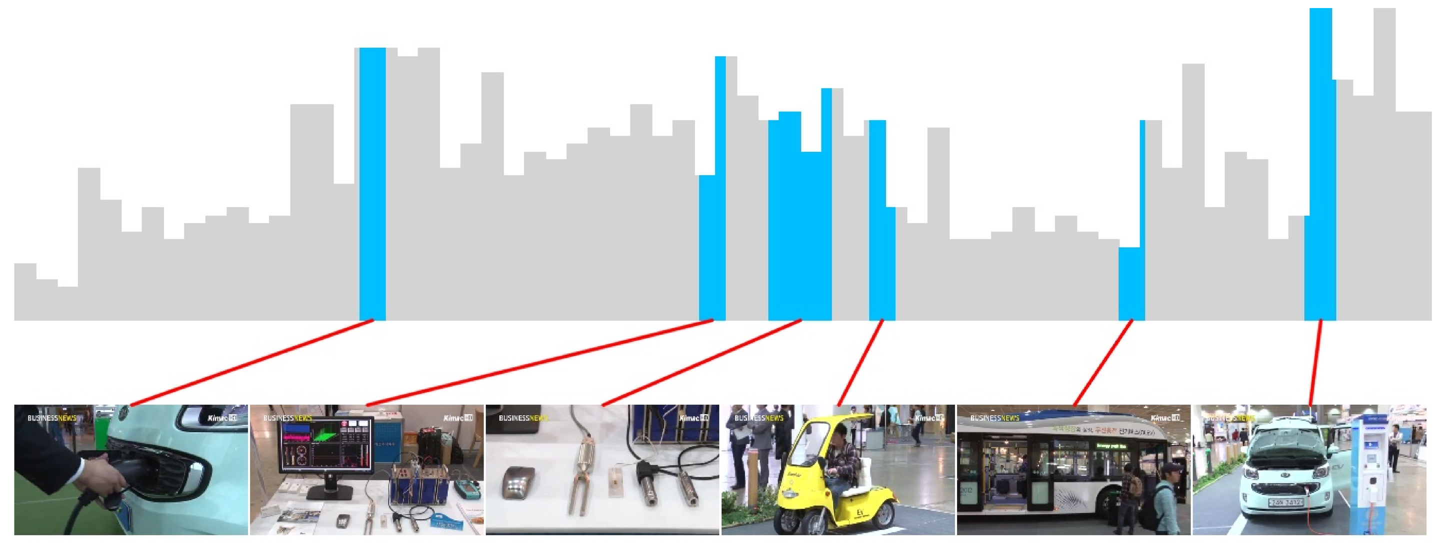
- Employes only a self-attention technique without RNN
- Using frame importance scores, self-attention relates each input sample to other input samples
- We use global attention where all input samples are considered at every step
- Video input is a sequence of CNN GoogLeNet features
- Network learns and predicts frame-level importance scores
- Frame scores are converted to binary keyshots summary limited to 15% of the original video length
- Model is trained by minimizing MSE loss with ADAM optimizer
- Easy to vectorize - entire video sequence can be processed in a single forward/backward operation without loops

Examples and Source Code

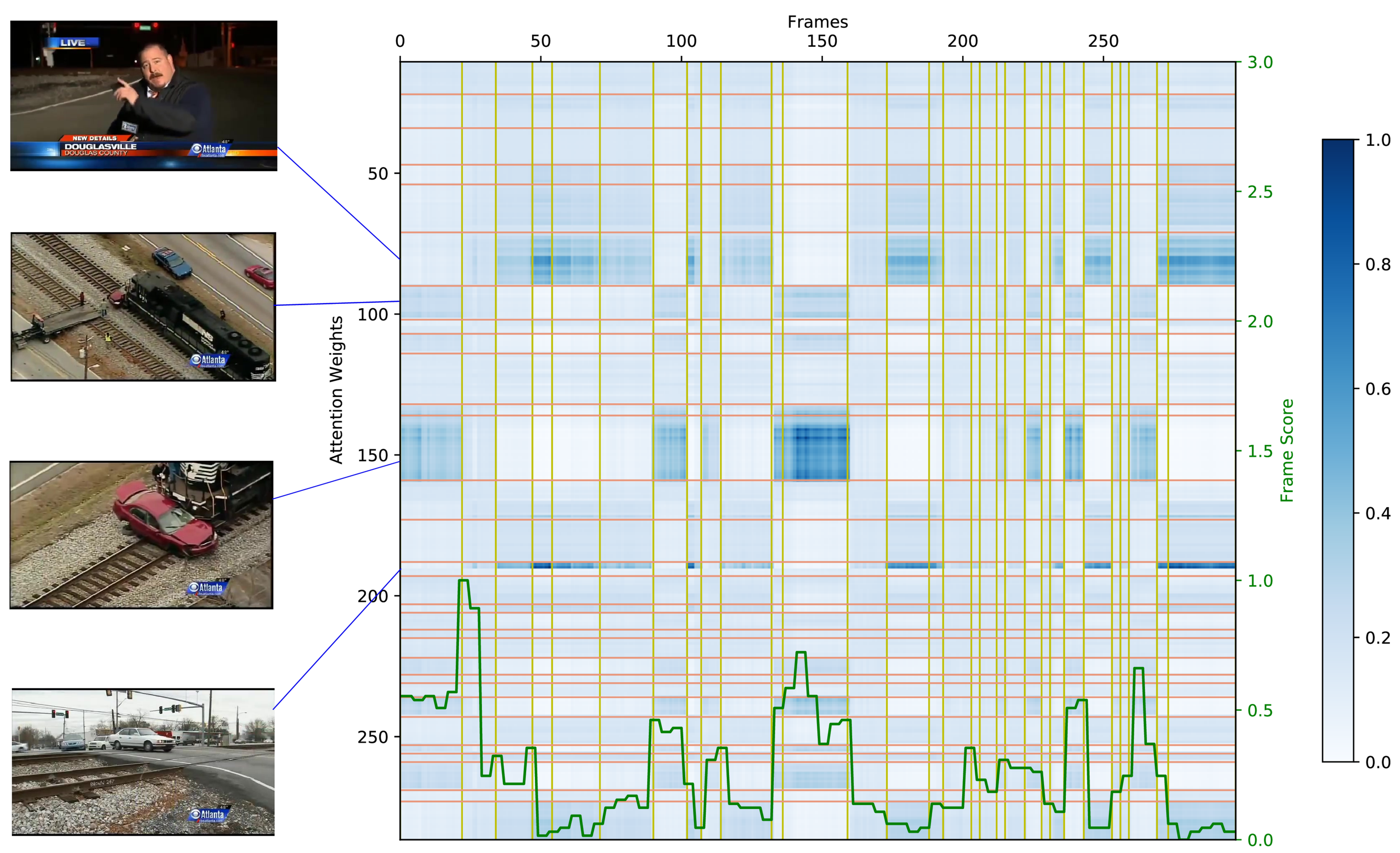
- Source Code (PyTorch) https://github.com/ok1zjf/vasnet/
- Examples https://goo.gl/cZkfJL

Qualitative Results





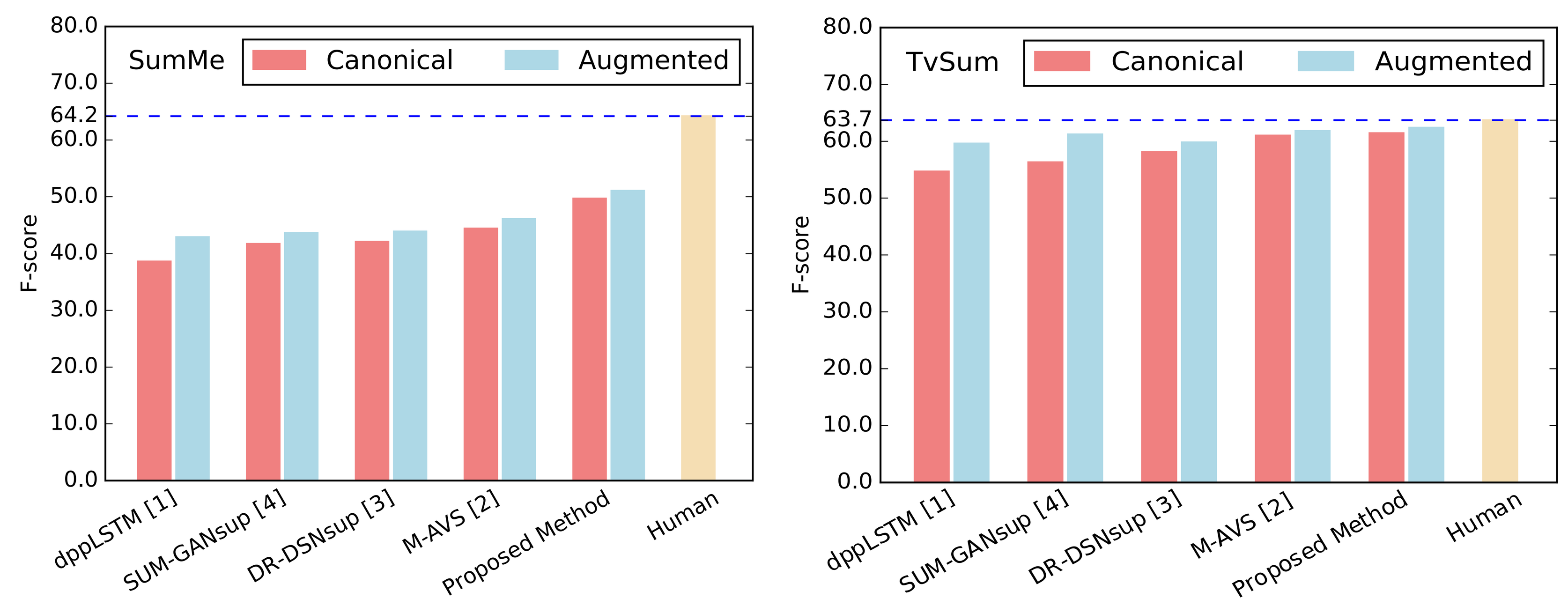
- Machine summary (blue) and ground truth (GT) (gray) for TvSum test video 10, split 4
- Selected keyshots align with most peaks in the GT and cover the entire video length
- Produces perceptually complete summary (please see examples)



- Confusion matrix of attention weights (blue) and GT frame scores (green) for TvSum test video 7 and shot boundaries (red/green, horizontal/vertical lines)
- The network learns to associate every video frame with other frames of similar score levels
- High gradient of the attention weights correlates with shot boundaries

Quantitative Results

Method	SumMe		TvSum	
	Canonical	Augmented	Canonical	Augmented
dppLSTM [1]	38.6	42.9	54.7	59.6
M-AVS [2]	44.4	46.1	61.0	61.8
DR-DSN$_{sup}$ [3]	42.1	43.9	58.1	59.8
SUM-GAN$_{sup}$ [4]	41.7	43.6	56.3	61.2
SASUM$_{sup}$ [5]	45.3	-	58.2	-
Human	64.2	-	63.7	-
VASNet (proposed method)	**49.71**	**51.09**	**61.42**	**62.37**



- Results reported as F-score in precentages
- Trained on TvSum, SumMe, YouTube and OVP datasets with canonical and augmented settings
- Evaluated on TvSum and SumMe over 5-fold splits
- Human performance is measured as an average F-score between training ground truth and all user summaries

References

[1] K. Zhang et al. Video summarization with long short-term memory. ECCV 2016, pp. 766–782.
[2] Z. Ji et al. Video summarization with attention-based encoder-decoder networks. arXiv preprint.
[3] K. Zhou et al. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. AAAI 2018.
[4] B. Mahasseni et al. Unsupervised video summarization with adversarial lstm networks. CVPR 2017, pp. 2982–2991
[5] H. Wei et al. Video summarization via semantic attended networks. AAAI 2018.

Acknowledgement

NATO OTAN The NATO Science for Peace and Security Programme
MONICA
NVIDIA.