Summarizing Videos with Attention

ACCV/AIU 2018

Jiri Fajtl[1], Hajar Sadeghi Sokeh[1], Vasileios Argyriou[1], Dorothy Monekosso[2] and Paolo Remagnino[1]

[1]Kingston University, London, UK
[2]Leeds Beckett University, Leeds, UK

December 19, 2018

Supported by:

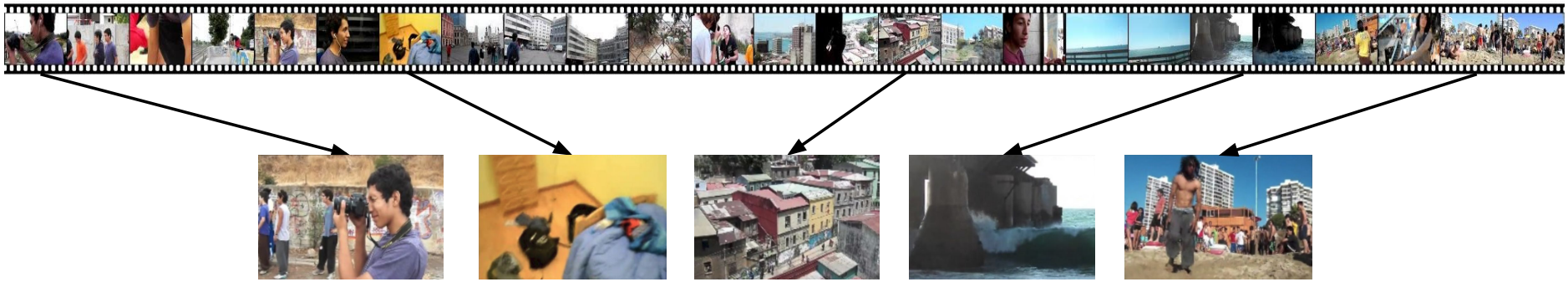
Video is Eating the Internet

- Amount of video data is huge and increasing
- Video will account for 80% of total Internet traffic in 2019
- There is a need for better tools to navigate through the video data

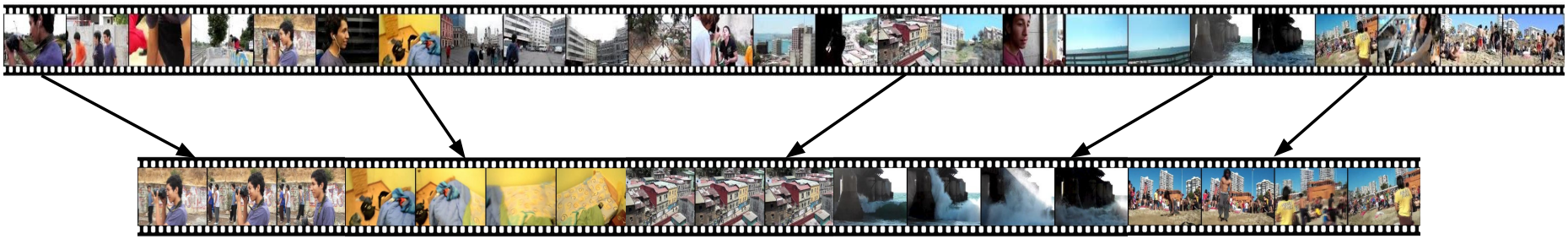
What is Video Summarization?

- Reduces video to a set of keyframes or keyshots
- The summary needs to preserve the most important information conveyed by the entire video
- Can be seen as a lossy "compression"

Keyframes Summarization (storyboard or thumbnails extraction)



Keyshots Summarization (skim or dynamic summaries)

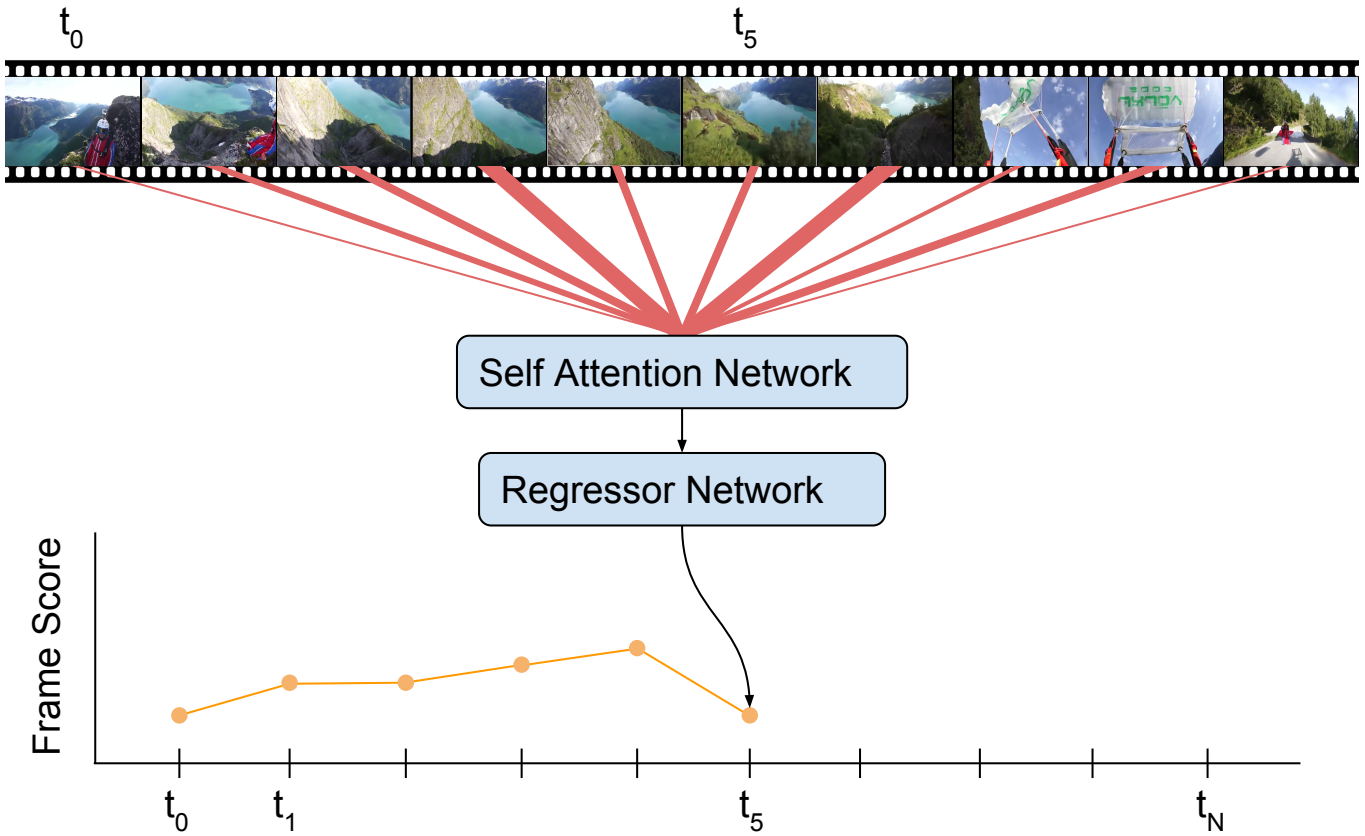


How to Summarize Videos

- Unsupervised methods try to devise features and cost measures based on assumptions about what makes a good summary
- Intuitively, video summarization is perceptually challenging task that requires deeper understanding of the video context
- Supervised methods "learn" from human annotations
- Deep architectures can leverage these annotations to learn complex relations within the video context to produce a human like summary
- Human summaries are subjective (low mutual agreement, F-score $\sim$ 33% on TvSum[1] and SumMe[2] datasets)
- We focus on a supervised, deep learning method for keyshots video summarization

Architecture

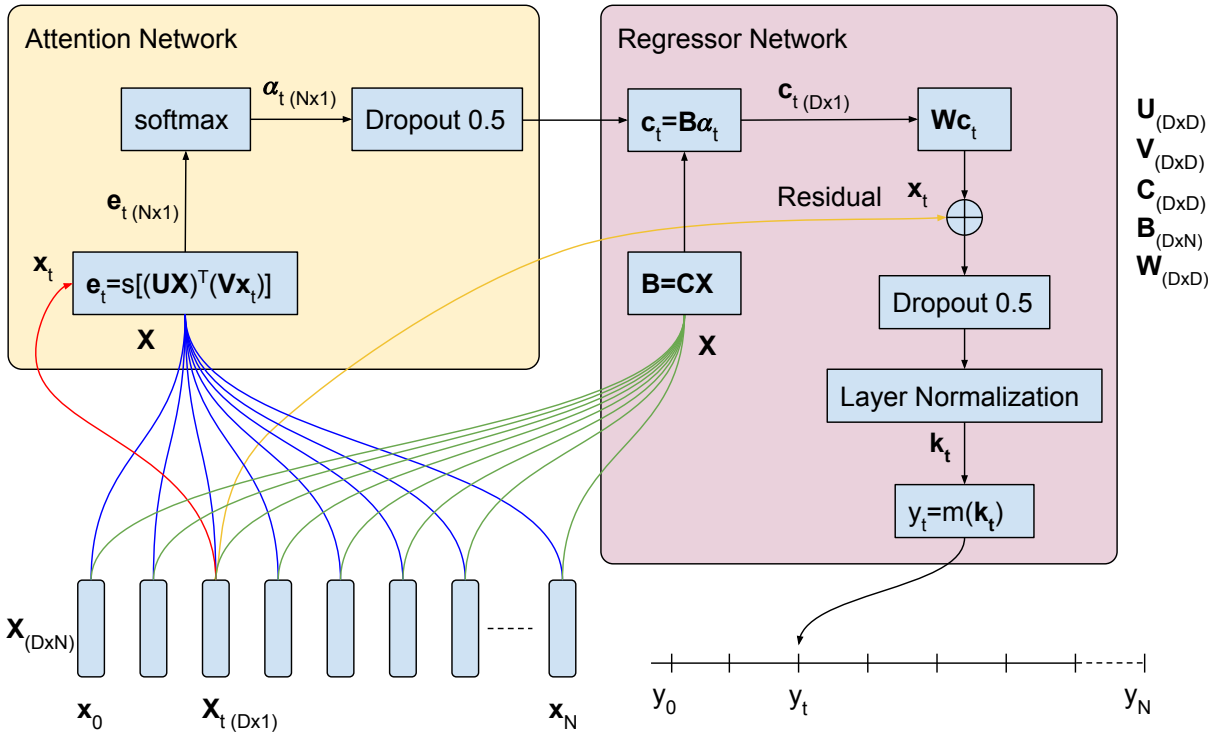
- Idea: Train a neural network with attention to learn the relation of each video frame with the rest of the sequence with respect to the user annotation



- Our model learns and predicts frame scores that are then converted to keyshots summary

Architecture

- Unlike state of the art methods we do not use RNN encoder-decoder but directly soft, self-attention
- Input are GoogLeNet CNN features $D = 1024 \times 1$ and output frame scores



α_t attention weights

$\mathbf{c}_t$ attention-weighted input sequence

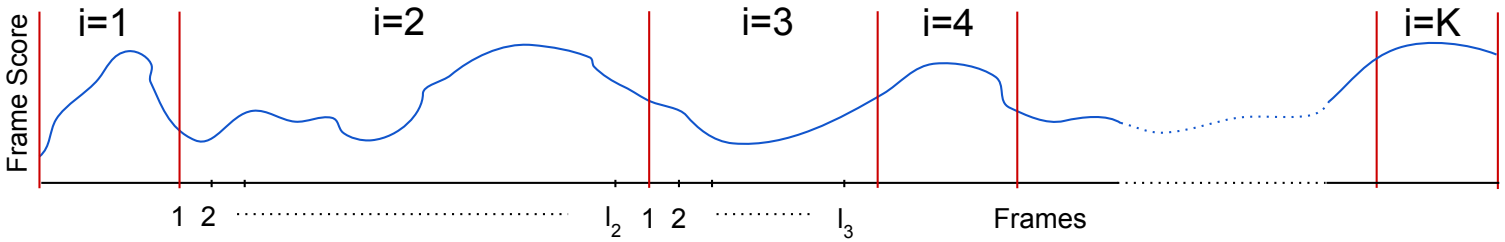
Training

- Model is trained by minimizing the MSE loss with the ADAM optimizer
- Trained up to 200 epochs
- Aggressive dropout regularization due to small training datasets
- Generalization attained by selecting models with highest validation accuracy
- Dataset annotations are converted to frame scores for training and to keyshot summaries for evaluation.

Dataset	**Videos**	**Annotation**	**Users**	**Mutual users F-score**	**GT vs users F-score**
TvSum	50	frame scores	20	53.8	63.7
SumMe	25	keyshots	15-18	31.1	64.2
YouTube	39	keyframes	5	-	-
OVP	50	keyframes	5	-	-

Inference

Partition video into K shots with KTS [3]. K is determined by KTS.



For each i-th shot we then calculate score s_i .

$$s_i = \frac{1}{l_i} \sum_{a=1}^{l_i} y_{i,a} \tag{1}$$

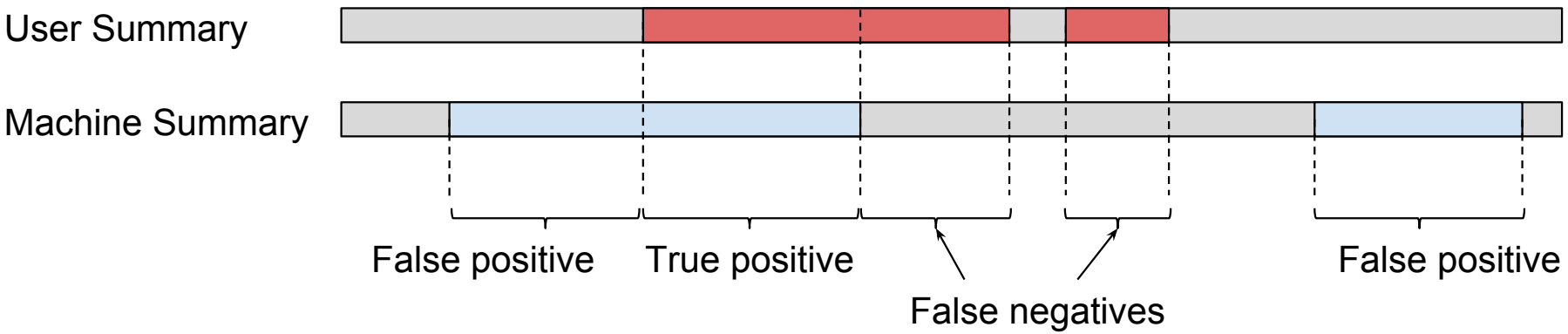
Where $y_{i,a}$ is score of a-th frame within shot i and l_i is length of i-th shot.

$$\max \sum_{i=1}^{K} u_i s_i, \quad \text{s. t.} \quad \sum_{i=1}^{K} u_i l_i \leq L, u_i \in 0, 1 \tag{2}$$

With L being a summary duration limit in frames (15% of original length). Summary is then a concatenation of all shots with $u_i = 1$

Evaluation

- We follow evaluation protocol according to [4],[2] and [1]
- All datasets used for training, TvSum and SumMe for evaluation
- 5-fold cross validation in canonical and augmented configurations
- Canonical: 80% of the eval dataset for training and 20% for test
- Augmented: all datasets + 80% of the eval dataset for training and 20% of the eval dataset for test
- Results reported as F-score in percentages



$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100 \tag{3}$$

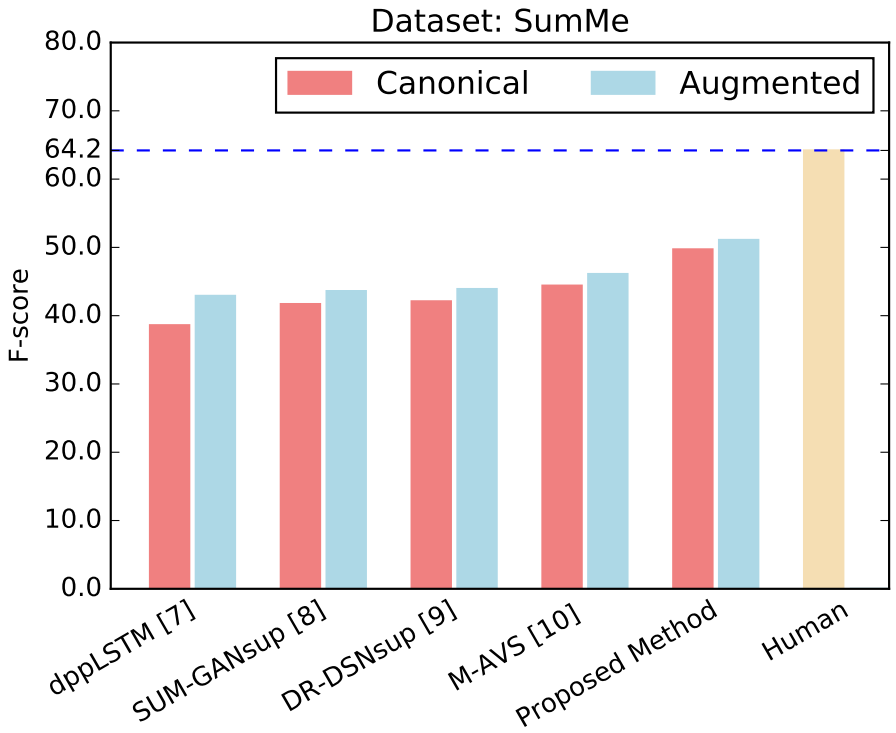
Results

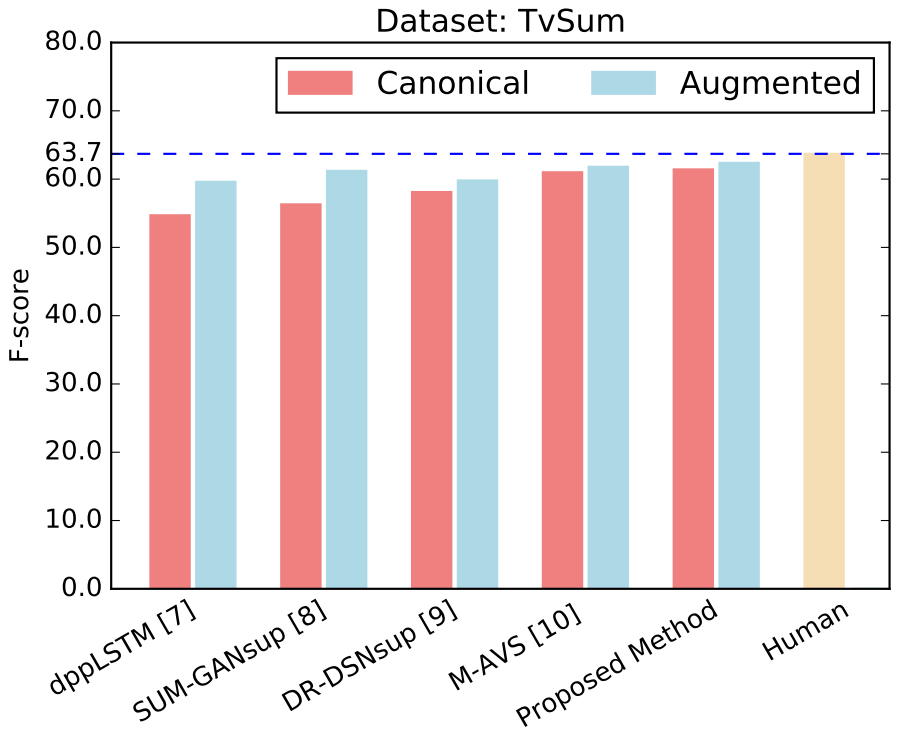
Table: Comparison of our method VASNet with the state of the art methods for canonical and augmented settings. For a reference we add human performance measured as an average F-score between training ground truth and all user summaries.

Method	SumMe		TvSum	
	Can	**Aug**	**Can**	**Aug**
dppLSTM [5]	38.6	42.9	54.7	59.6
M-AVS [6]	44.4	46.1	61.0	61.8
DR-DSN$_{sup}$ [7]	42.1	43.9	58.1	59.8
SUM-GAN$_{sup}$ [8]	41.7	43.6	56.3	61.2
SASUM$_{sup}$ [9]	45.3	-	58.2	-
Human	64.2	-	63.7	-
VASNet (proposed method)	**49.71**	**51.09**	**61.42**	**62.37**

Results

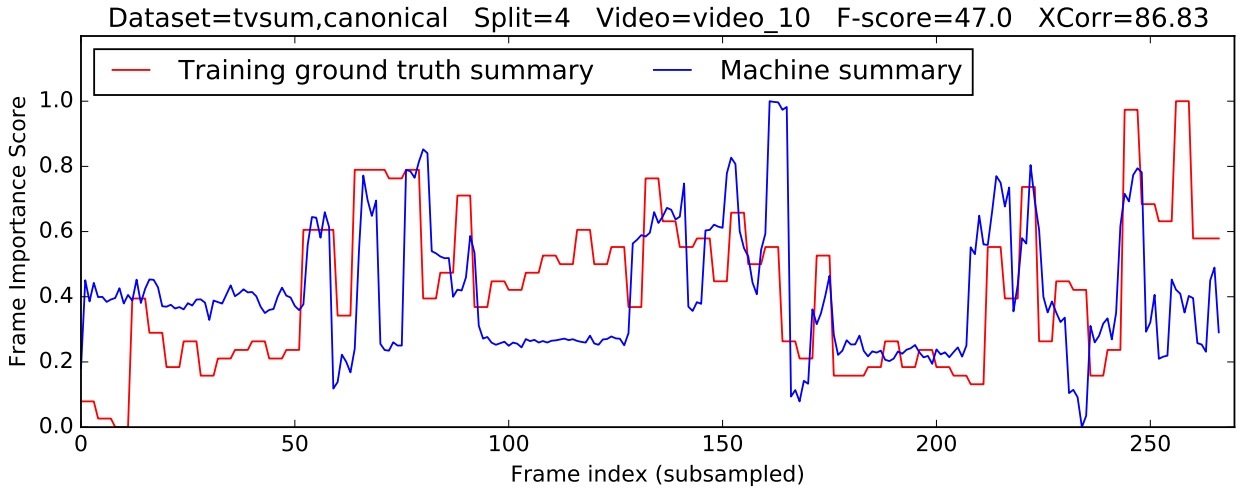
Our method outperforms current state of the art, supervised methods for video summarization in all settings and on both datasets.

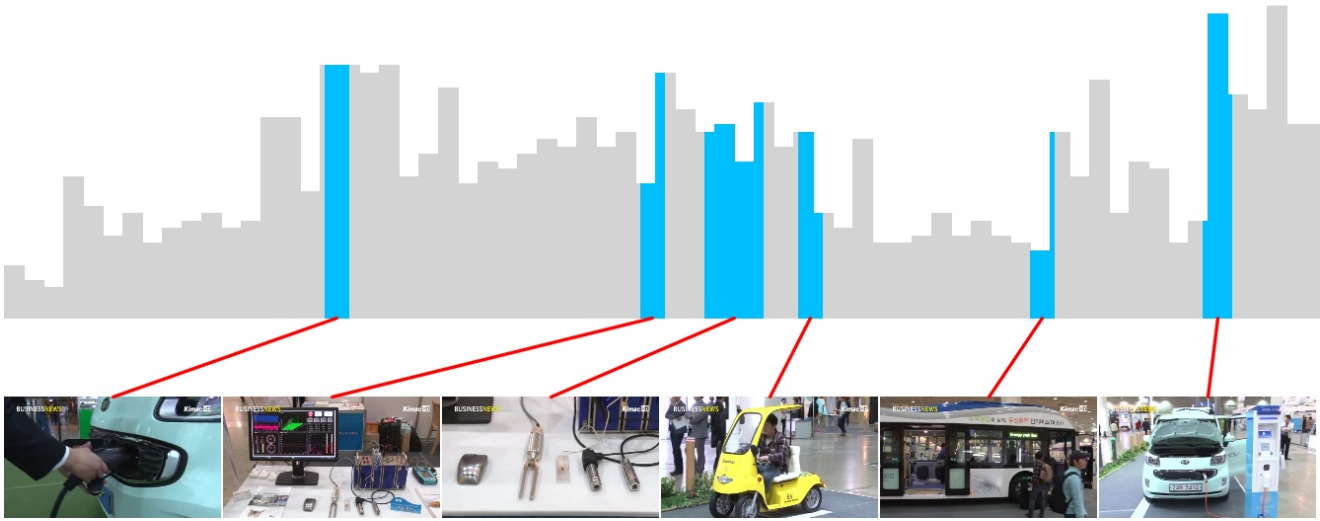




Qualitative Results

Machine summary and ground truth for TvSum video 10, test split 4.

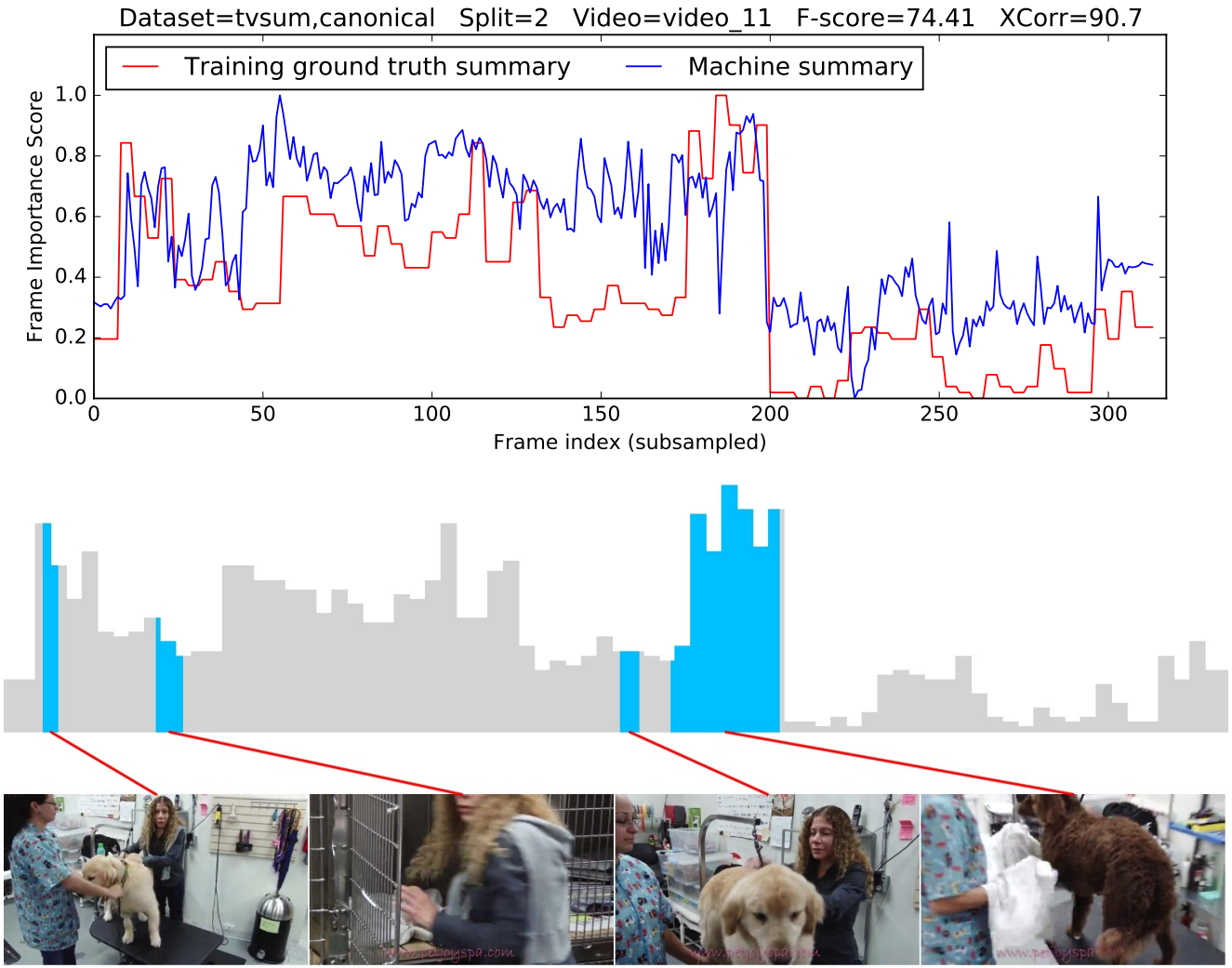




Summary video: `https://youtu.be/m8jPXan1phc`

Qualitative Results

Machine summary and ground truth for TvSum video 11, test split 2.

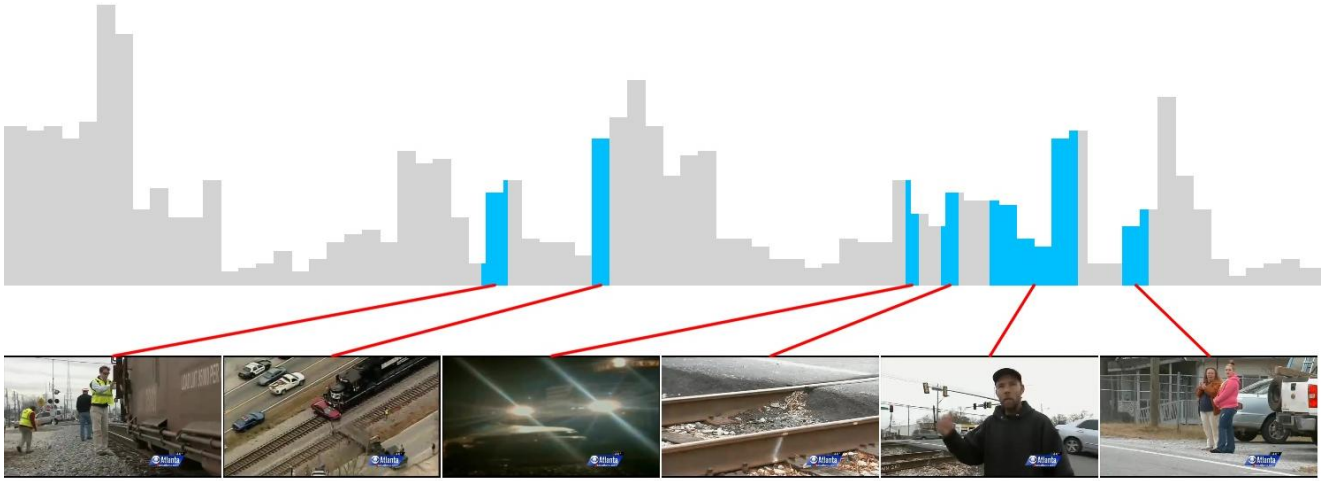


Summary video: `https://youtu.be/ZdAQLHBxDtc`

Qualitative Results

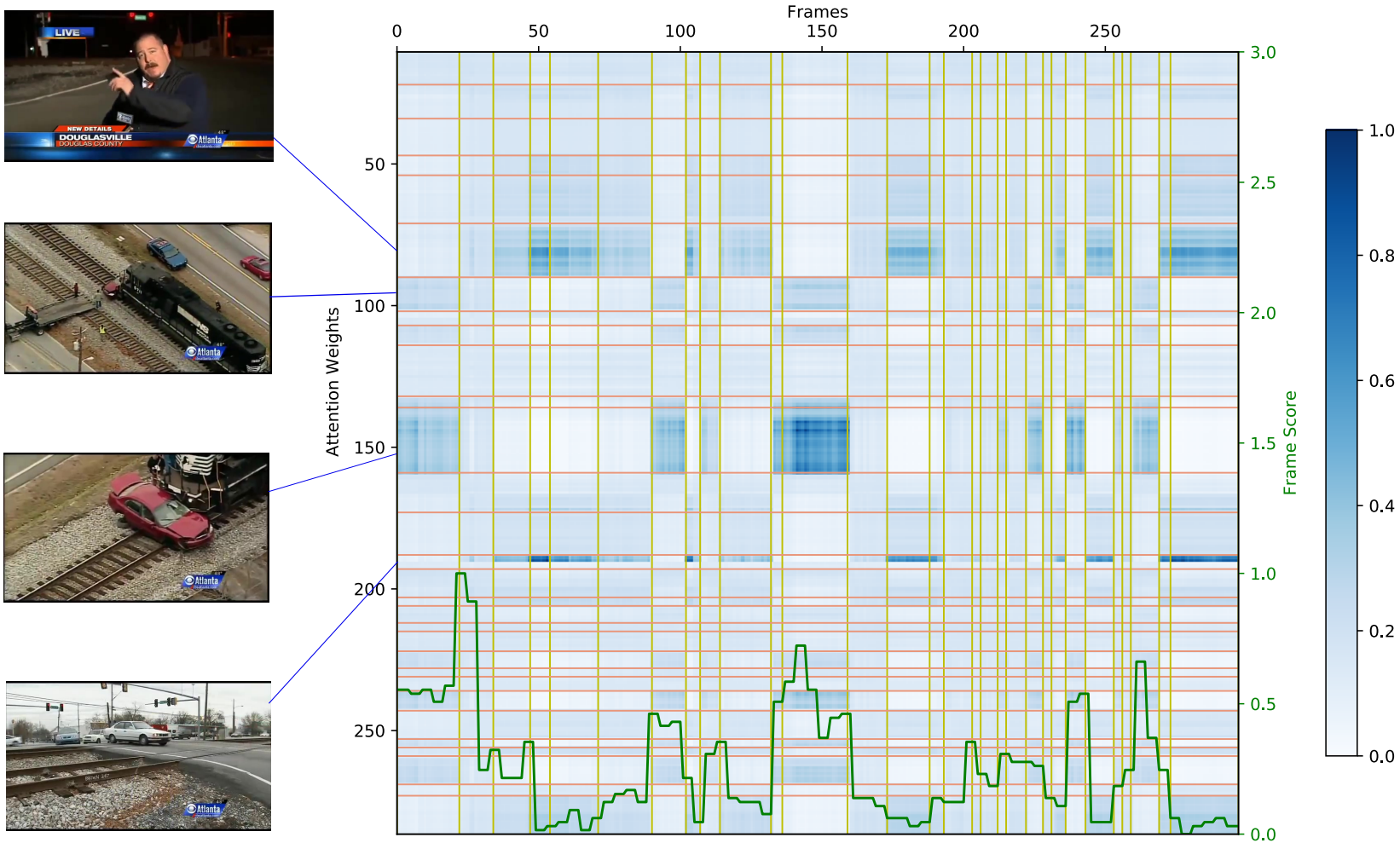
- Confusion matrix of the attention weights (next slide) offers an insight into what the model learns
- The attention clusters scene shots correlated with high and low frame scores (green plot at the bottom of the matrix)
- Also indirectly learns shot boundaries (red/green matrix grid)

Machine summary for TvSum video 7, test split 3.



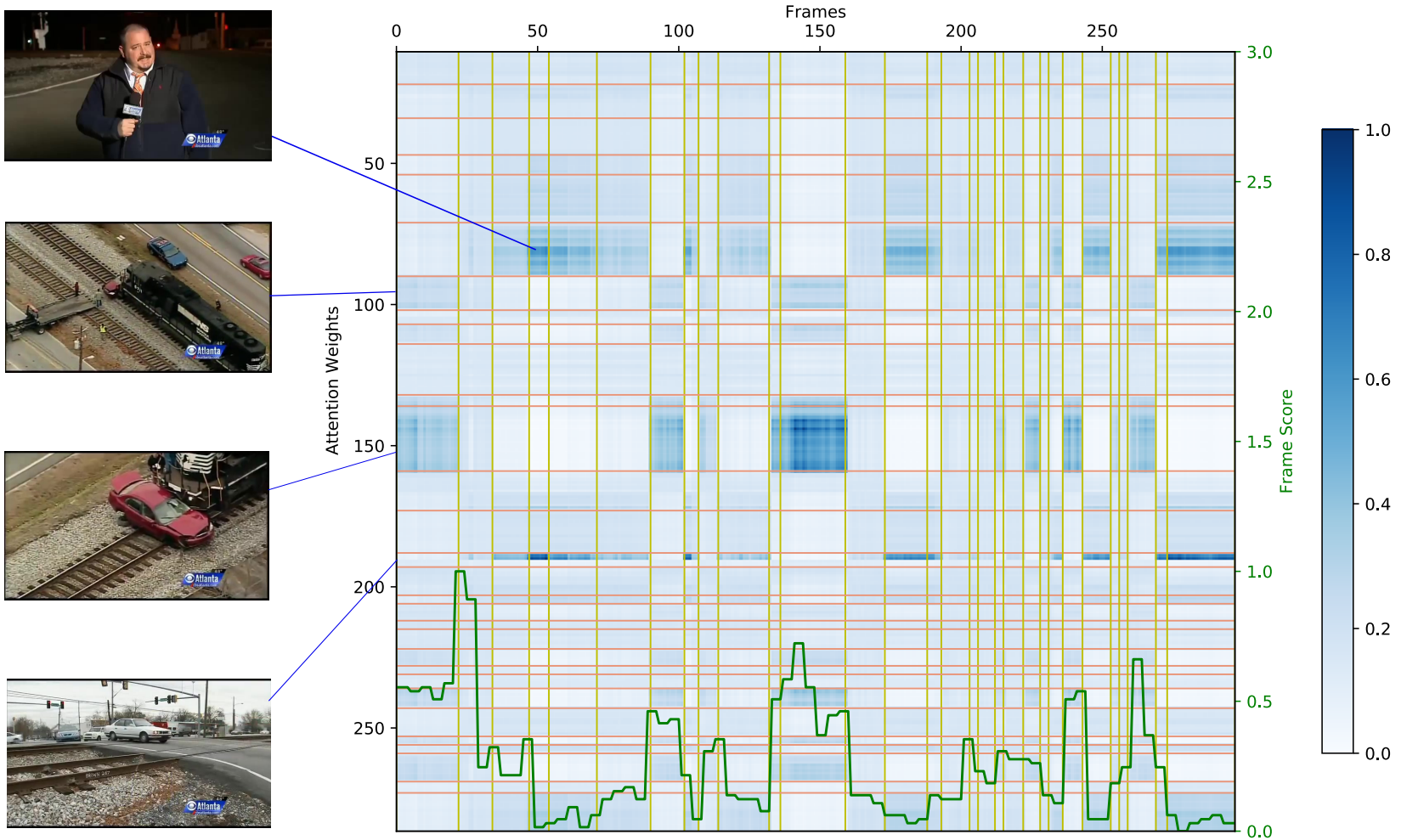
Summary video: `https://youtu.be/bCcCQ-qsjpg` []

Qualitative Results



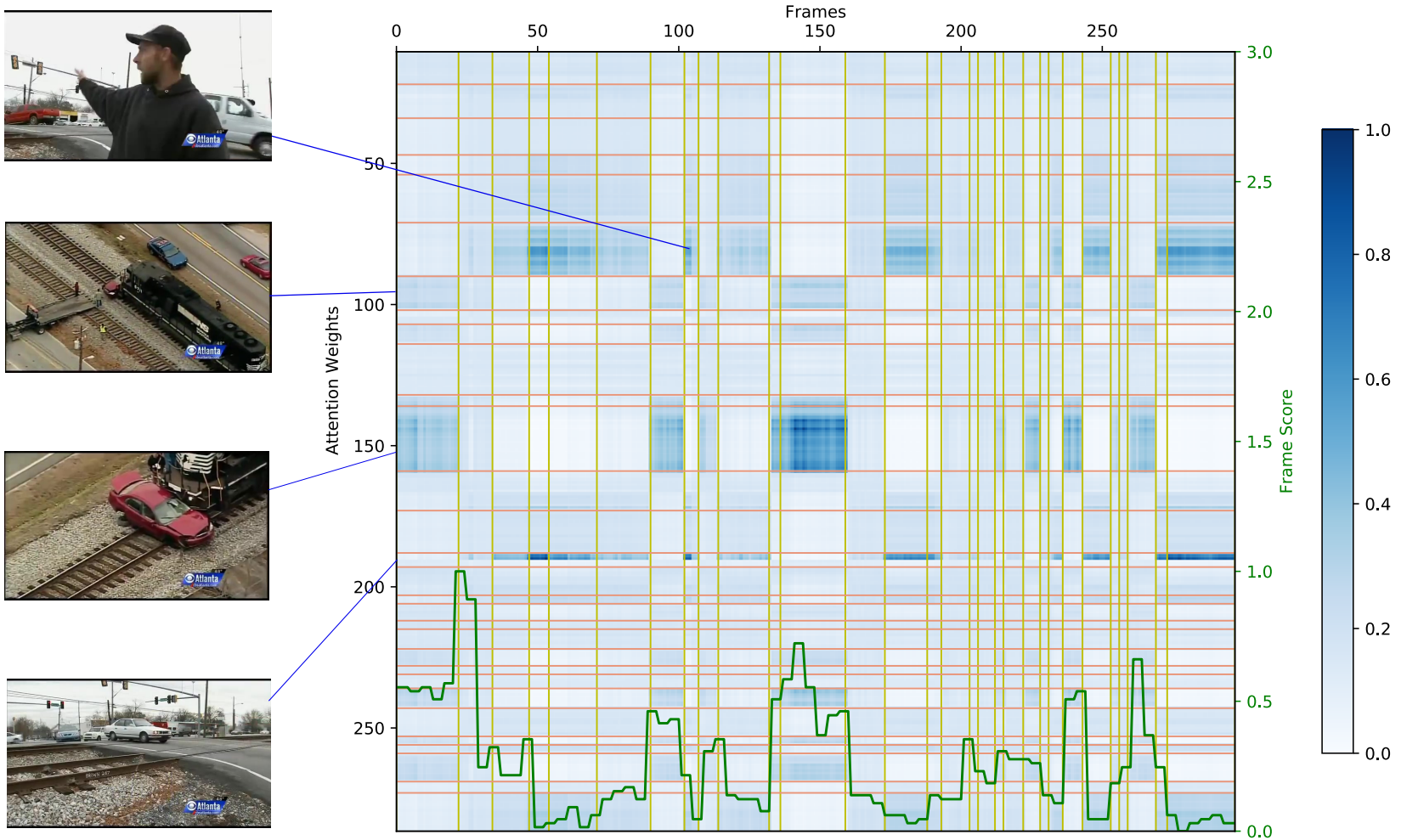
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



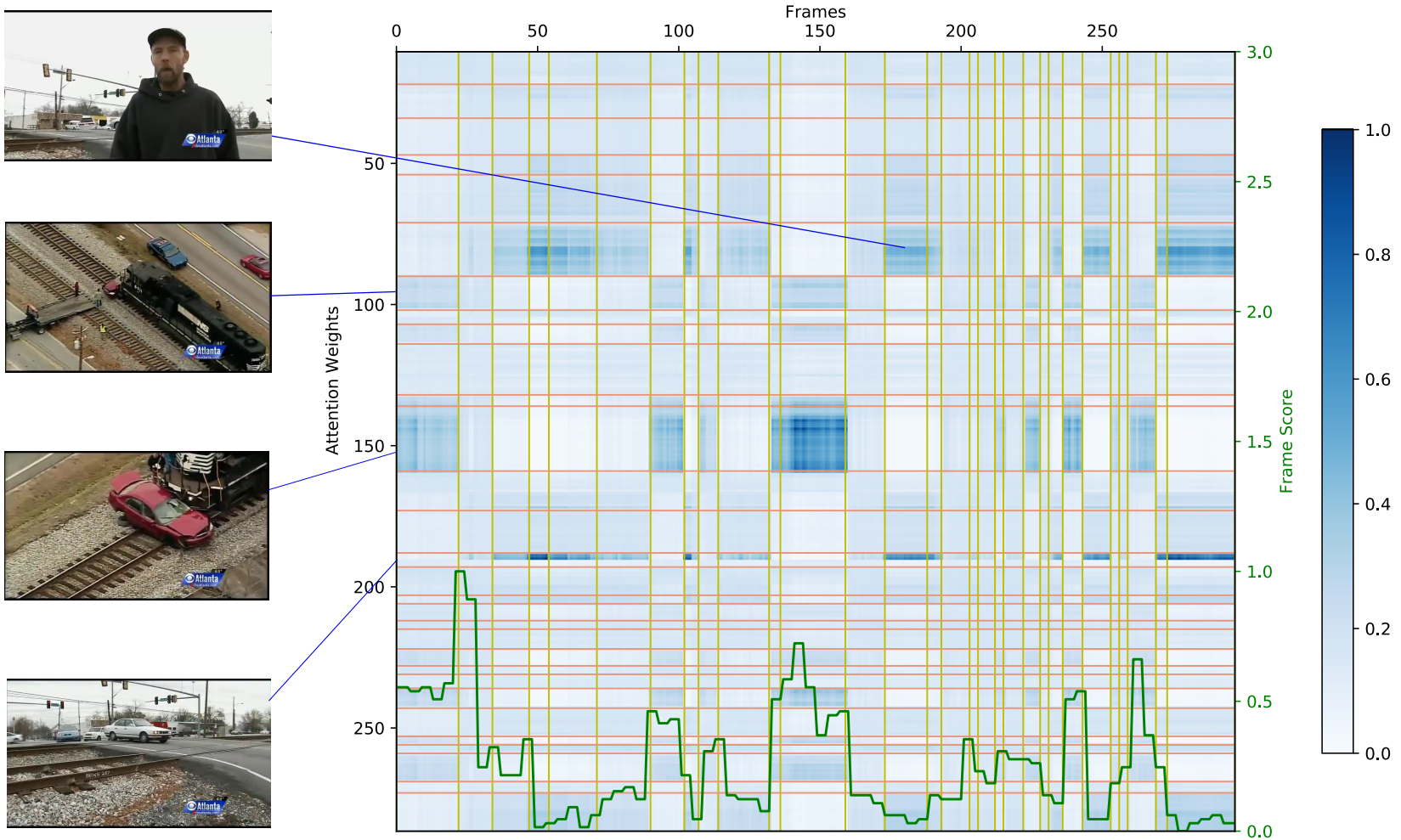
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



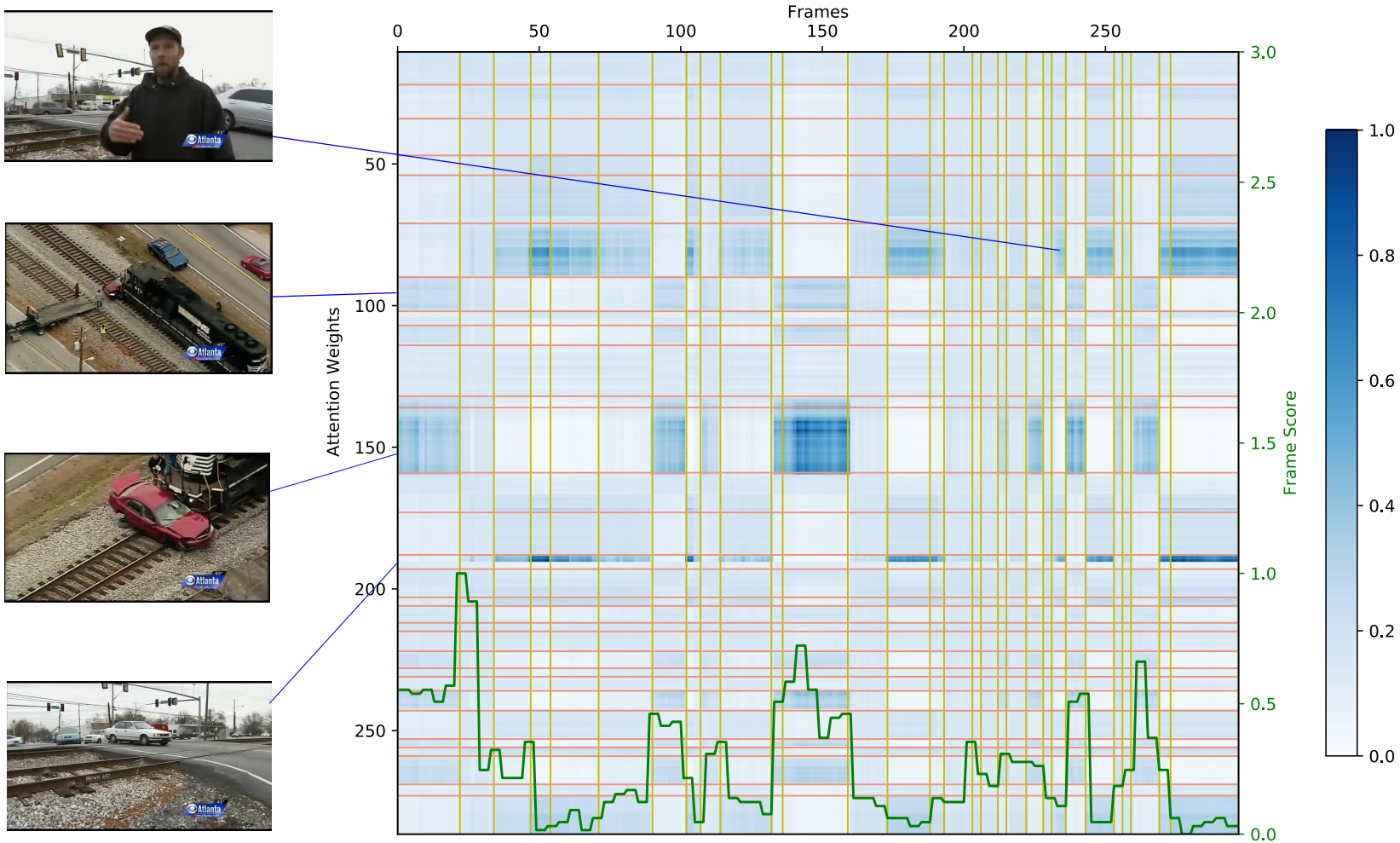
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



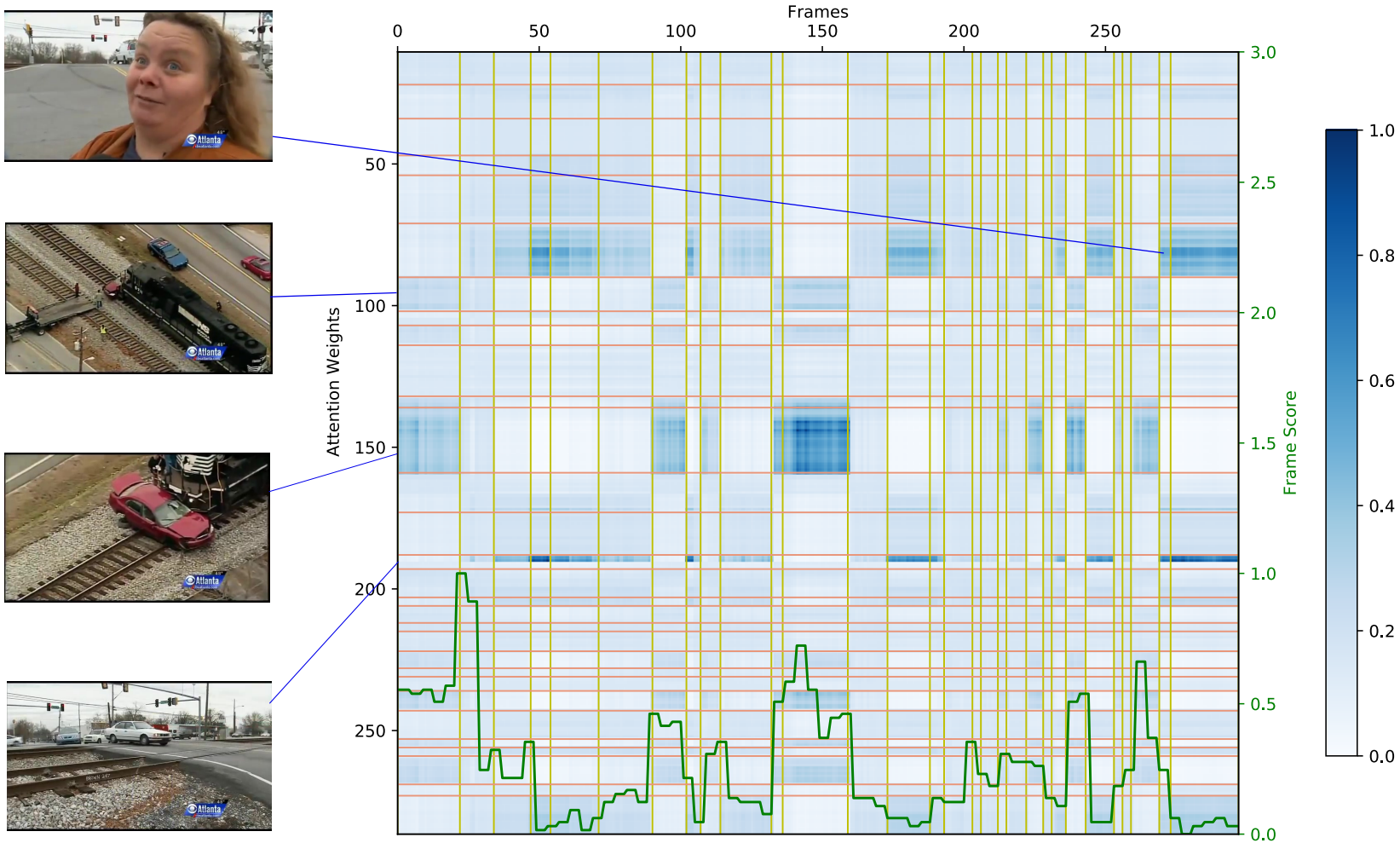
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



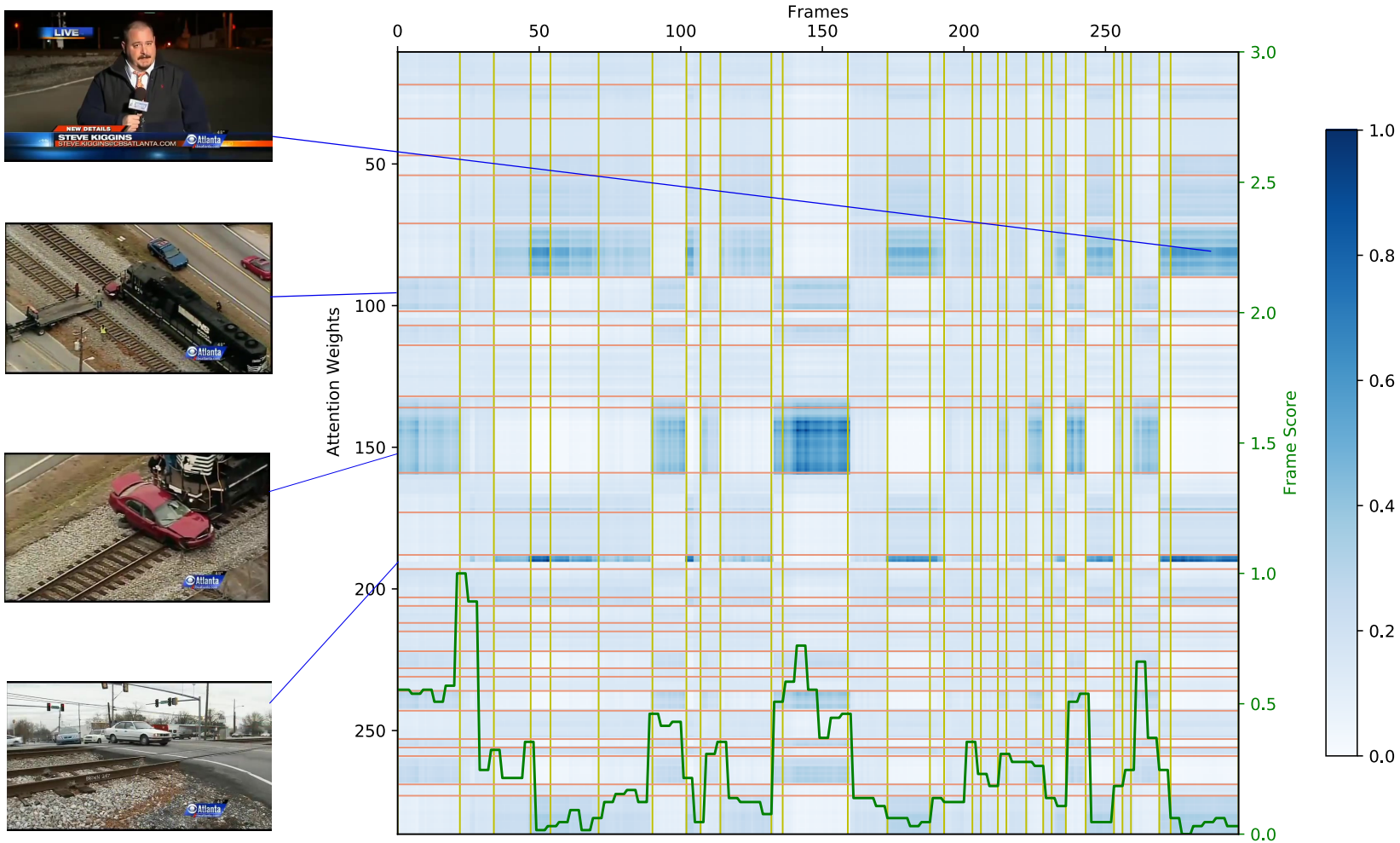
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



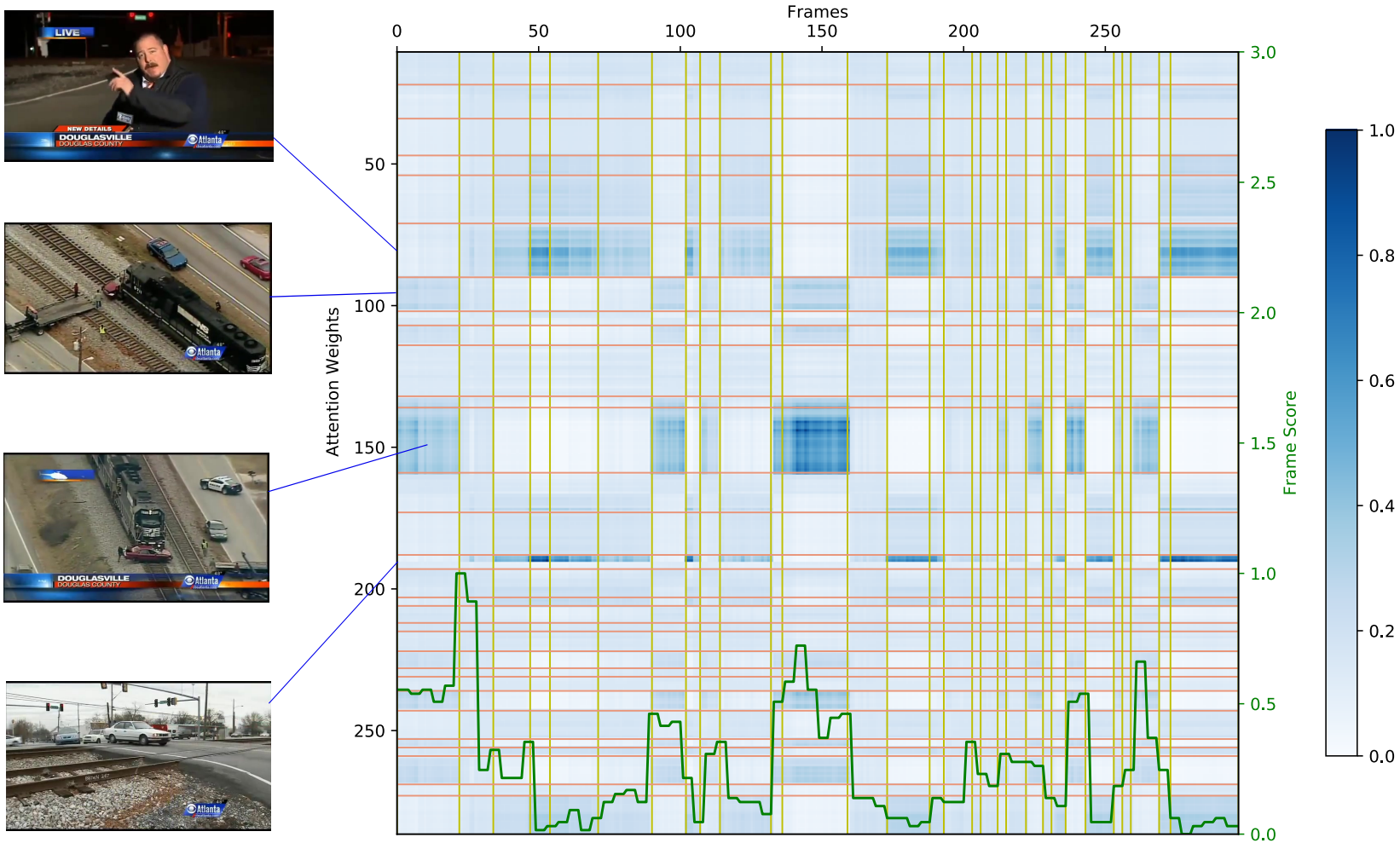
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



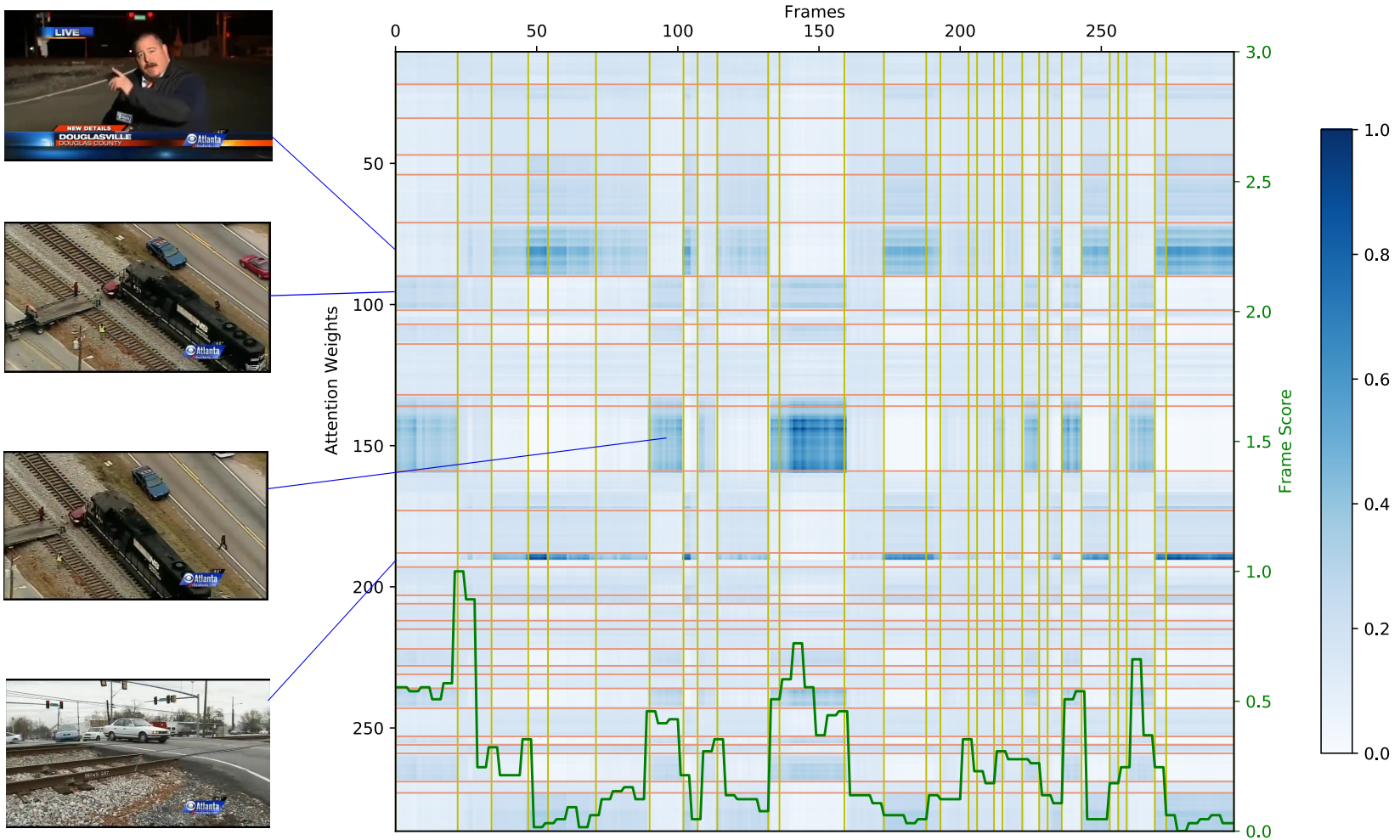
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



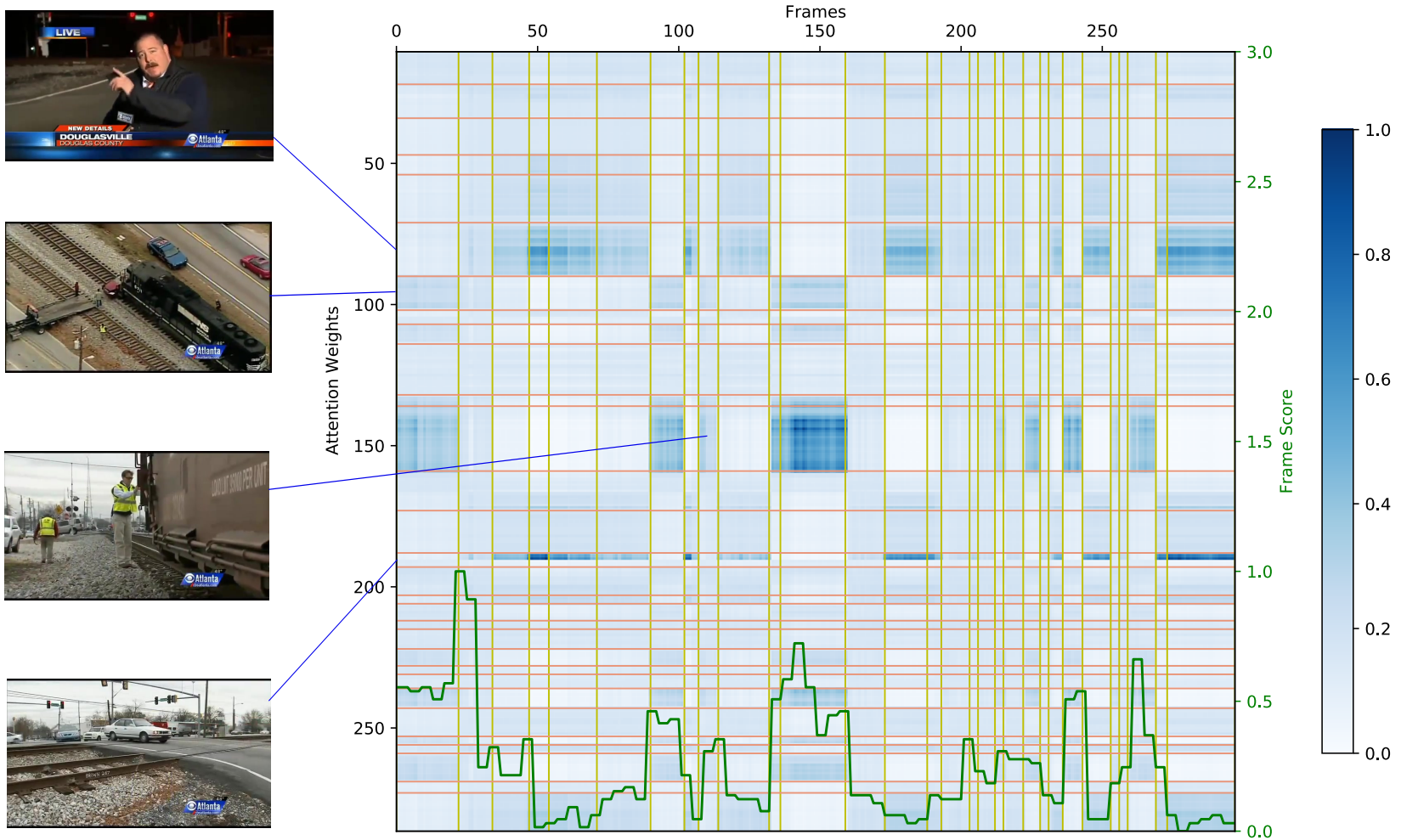
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



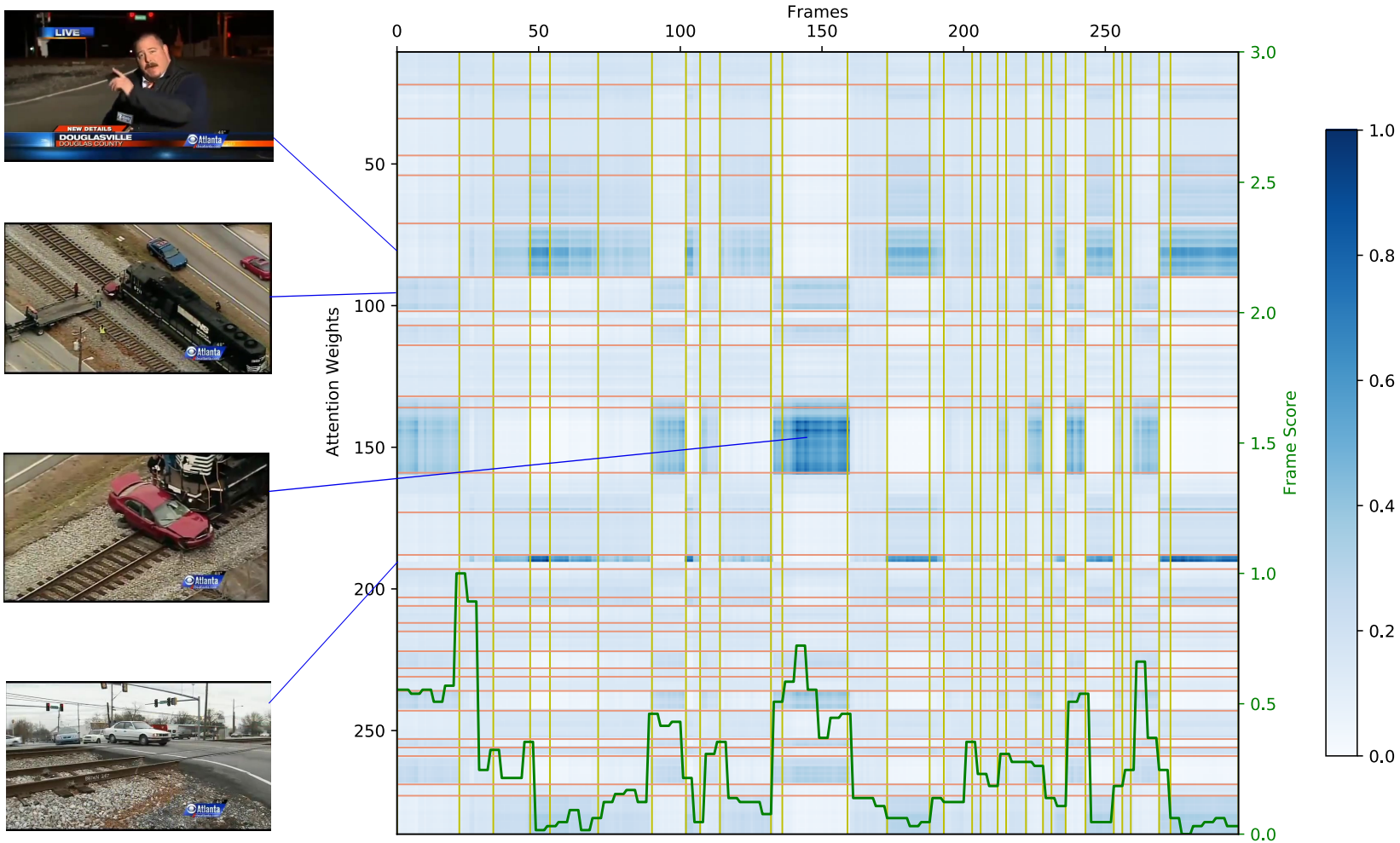
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



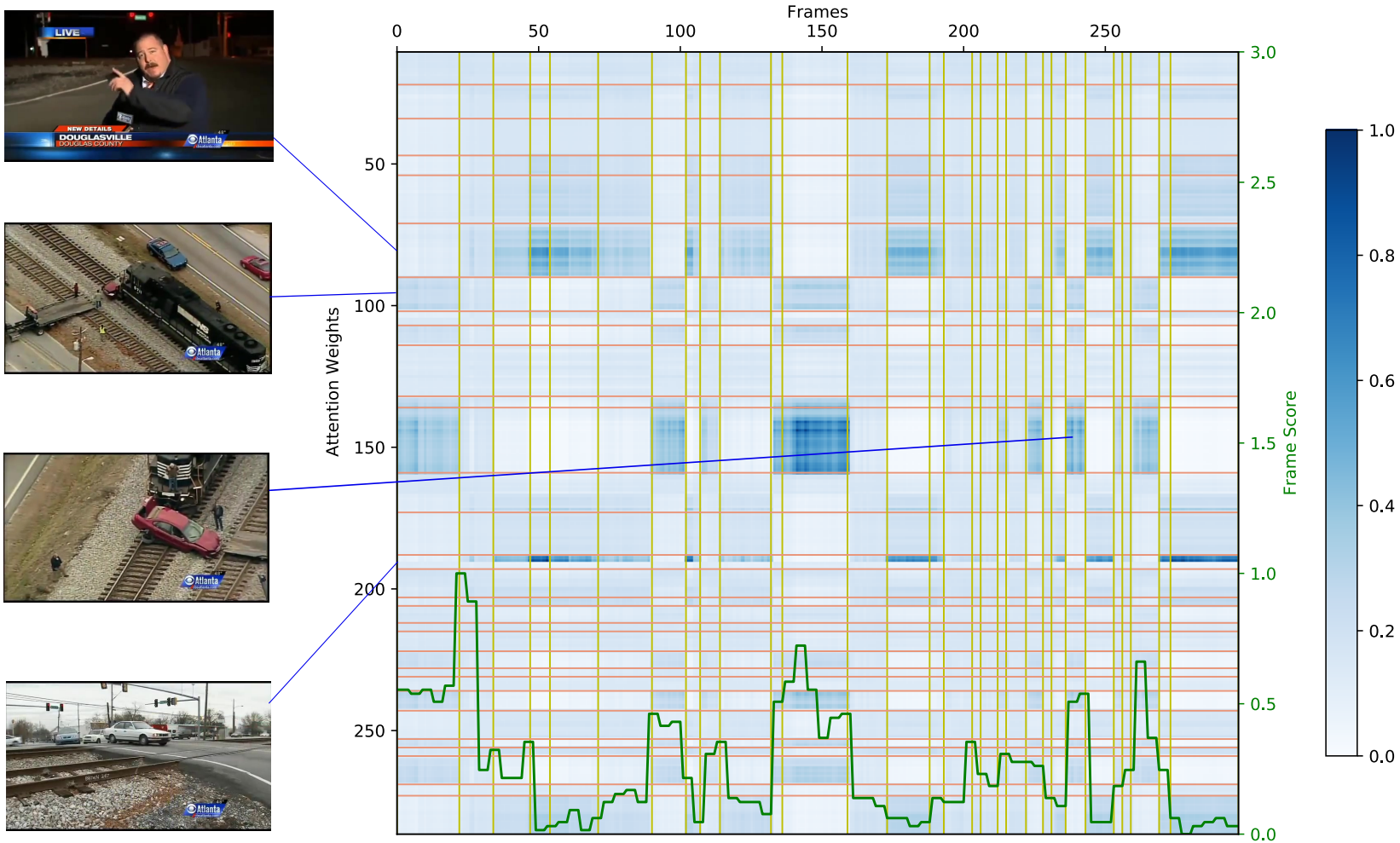
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



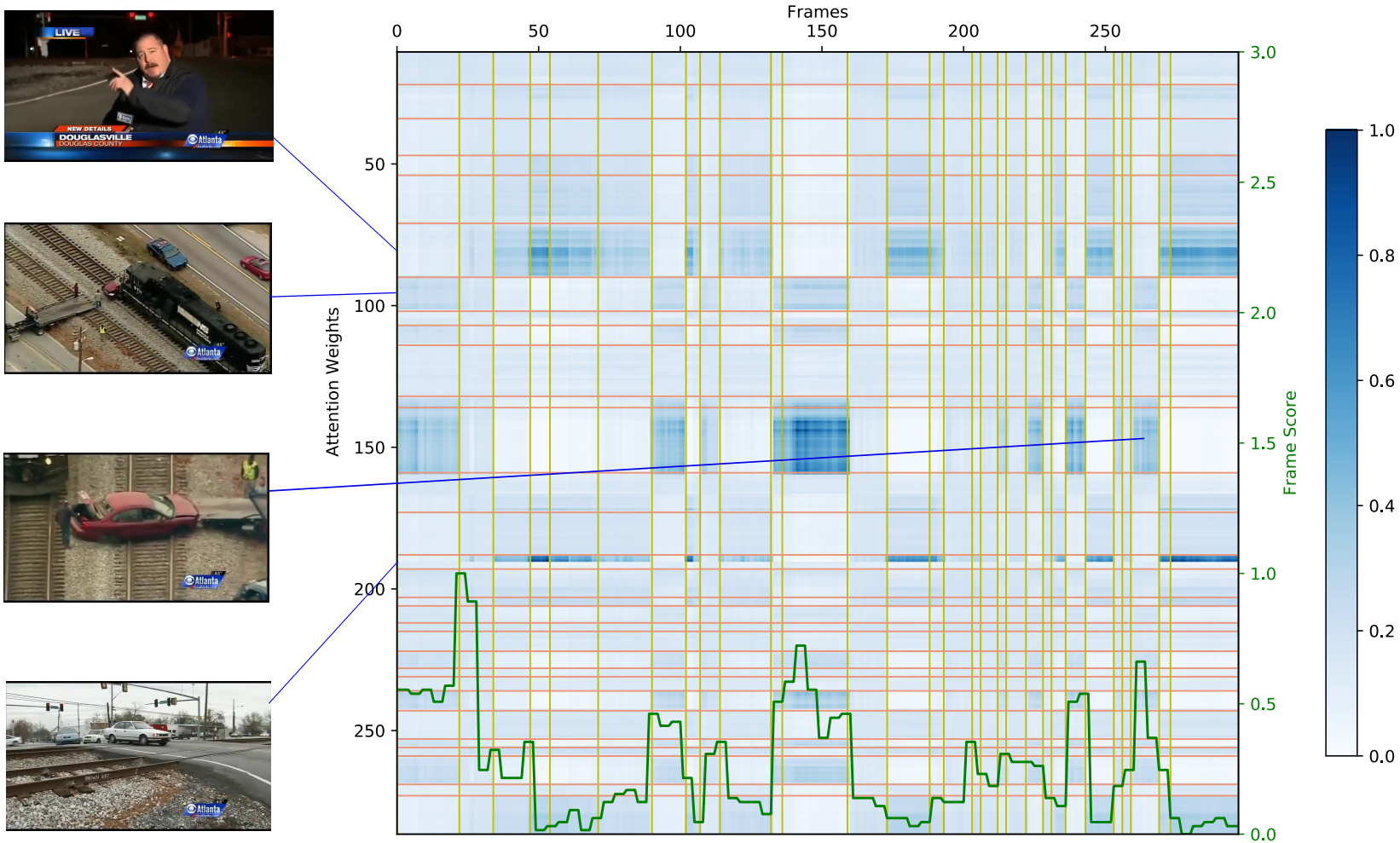
Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Qualitative Results



Confusion matrix of attention weights (blue) and ground truth frame scores (green) for TvSum video 7, test split 3.

Conclusion

- Presented soft, self-attention model for sequence to sequence processing
- Showed that the attention-only method outperforms complex encoder-decoder architectures on supervised video summarization task

Future Work

- Implement and evaluate self-attention with local aperture
- Train and evaluate VASNet on longer first/third person view videos
- Train the model to directly estimate keyshot boundaries and keyshot scores rather than the frame scores - remove dependency on KTS

Code and Examples

- **Source code and trained models** (PyTorch 0.4)
  `https://github.com/ok1zjf/vasnet/`

- **Examples of summarized videos**
  `https://www.youtube.com/playlist?list=PLEdpjt8KmmQMfQEat4HvuIxORwiO9q9DB`

- **Contact**
  `mailto:J.Fajtl@kingston.ac.uk`

Thank You!

References I

[1] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE CVPR*, pp. 5179–5187, 2015.

[2] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proceedings of the ECCV*, pp. 505–520, Springer, 2014.

[3] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proceedings of the ECCV*, pp. 540–555, Springer, 2014.

[4] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the IEEE CVPR*, pp. 3090–3098, 2015.

References II

[5] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proceedings of the ECCV*, pp. 766–782, Springer, 2016.

[6] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *arXiv preprint arXiv:1708.09545*, 2017.

[7] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the AAAI*, 2018.

[8] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," *Proceedings of the IEEE CVPR*, pp. 2982–2991, 2017.

References III

[9] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Proceedings of the AAAI*, 2018.