Video2GIF: Automatic Generation of Animated GIFs from Video - Michael Gygli, Yale Song, Liangliang Cao (CVPR 2016)

Gaurav Kakkar Department of CSE, IIT Kanpur
Palash Chauhan Department of CSE, IIT Kanpur
Shubham Agarwal Department of CSE, IIT Kanpur

This paper introduces the novel problem of generating animated GIFs from videos. GIFs are short looping videos with no sound, and a perfect combination between image and video that really capture our attention by expressing various forms of emotions. The task has connections to problems like video highlights, visual interestingness and summarization and applications in areas like photojournalism and advertisements. This problem has not been addressed before, however it is closely related to the work by *Sun et al*[35] and *Potapov et al*[30] who propose domain-specific models to predict video highlights. Various websites like GIFSoup, Imgflip, and Ezgif provide tools to generate GIFs from videos which are cumbersome and require extensive human effort since the user specifies the exact time range in the video for the creation of the GIF.

Using the websites mentioned above, a large-scale dataset was collected consisting of over 120K animated GIFs extracted from more than 80K videos. Each GIF was aligned to its video using frame matching. A perceptual hash based on the discrete cosine transform[41] was used to encode every frame and matching was done using the Hamming distance. This alignment was essential since the non-selected segments serve as negative training samples. Videos with duration more than 10 minutes were discarded since GIF segments become too sparse and the videos are more affected by chronological bias [33]. The final training, validation and test set consist of 65K, 5K and 357 videos respectively.

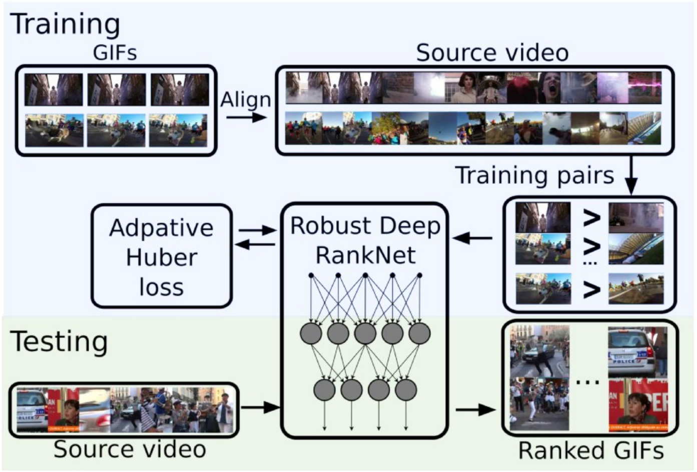

Figure 1: Workflow

Rather than posing as a classification problem, paper formulates it as a ranking problem. Video is divided into non-overlapping segments using shot boundary detection algorithm *Song et al.*[33]. These segments may not be perfectly aligned with the actual GIF segments, so segment(s) with more than 66% overlap is considered as positive GIF segment(s^+).Figure 1 depicts the architecture of the model. During the learning phase,GIF and non GIF segments are the input to the model and it tries to learn a function $h(s) : \mathbb{R}^d \to \mathbb{R}$ which is defined as GIF-*suitability* score of a segment. Apart from the spatial and temporal features of the segement extracted using C3D[36], optional contextual features like category label, semantic embedding of the video tags and positional features are included. Rank constraints(1) over the dataset is formulated to be video-specific (S) as segment suitability across videos is not meaningful.

$$h(s^+) > h(s^-) \quad \forall (s^+, s^-) \in S \tag{1}$$

A novel adaptive Huber loss is presented in imposing the rank constraints. The classical l_1 and l_2 loss constraints require the positive segment to score higher than negative segment by margin of 1, but they suffer from the serious drawback of over-penalizing small margin violation and outliers respectively.

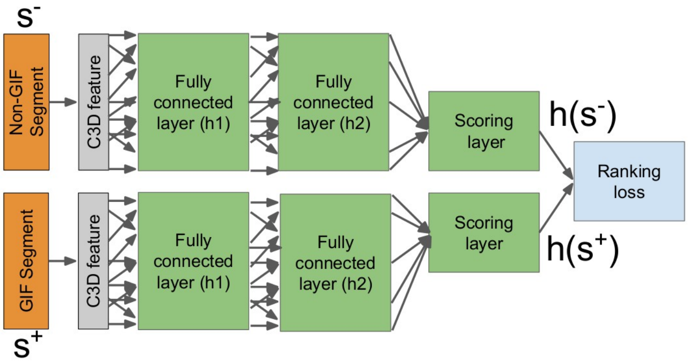
S-
Non-GIF Segment
C3D feature
Fully connected layer (h1)
Fully connected layer (h2)
Scoring layer
h(s-)
Ranking loss
GIF Segment
C3D feature
Fully connected layer (h1)
Fully connected layer (h2)
Scoring layer
h(s+)
S+

Figure 2: The architecture of the Robust Deep RankNet

$$l_{Huber}(s^+, s^-) = \begin{cases} \frac{1}{2} l_2(s^+, s^-), & ifu \le \delta \\ \delta l_1(s^+, s^-) - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{2}$$

Loss being squared for small margin violation and linear for outliers(stronger violation) better constrains ranking problem. GIFs cannot be assumed to be of equal quality, thus by introducing parameter δ ,which is GIF dependent, paper ensures a adaptive scoring scheme.

For evaluation, classical performance metrics - mean average precision(mAP)[35] and average meaningful summary detection(MSD)[30] - suffered from drawback of being sensitive to video length. A normalised nMSD is proposed which incorporates relative length of selected GIF at a recall rate of α

$$nMSD = \frac{|G^*| - \alpha|G^{gt}|}{|V| - \alpha|G^{gt}|} \tag{3}$$

where $|.|$ denots the length of a GIF or video and $|G^*|$ is the GIF with α recall w.r.t. the ground truth GIF G^{gt}.

Table 1 shows that the proposed method outperforms the baseline Deep visual-semantic embedding[27], Domain-specific rankSVM[35], and Category-specific summarization[30] leading to better results in terms of nMSD and mAP due to non linear neural network and novel huber loss robust to outliers.

Method	nMSD ↓	mAP ↑
Joint embedding [27]	54.38%	12.36%
Category-spec. SVM [30]	52.98%	13.46%
Domain-spec. rankSVM [35]	46.40%	16.08%
Classification	61.37%	11.78%
Rank, video agnostic	53.71%	13.25%
Rank, l_1 loss	44.60%	16.09%
Rank, l_2 loss	44.43%	16.10%
Rank, Huber loss	44.43%	**16.22%**
Rank, adaptive Huber loss	44.58%	16.21%
Rank, adaptive Huber loss + context (Ours)	44.19%	16.18%
Ours + model averaging	**44.08%**	16.21%
Approx. bounds	38.77%	21.30%

Table 1: Experimental results: A lower nMSD and higher mAP represents better performance

In conclusion, paper has proposed a Robust Deep RankNet that predicts the GIF suitability of video segments with efficient handling of noise and low quality web data with a novel adaptive Huber rank loss.