HUCVL at MediaEval 2016: Predicting Interesting Key Frames with Deep Models

Goksu Erdogan, Aykut Erdem, Erkut Erdem
Hacettepe Computer Vision Lab (HUCVL)
Department of Computer Engineering, Hacettepe University, Ankara, Turkey
{goksuerdogan, aykut, erkut}@cs.hacettepe.edu.tr

ABSTRACT

In MediaEval 2016, we focus on the image interestingness subtask which involves predicting interesting key frames of a video in the form of a movie trailer. We specifically propose three different deep models for this subtask. The first two models are based on fine-tuning two pretrained models, namely AlexNet and MemNet, where we cast the interestingness prediction as a regression problem. Our third deep model, on the other hand, depends on a triplet network which is comprised of three instances of the same feedforward network with shared weights, and trained according to a triplet ranking loss. Our experiments demonstrate that all these models provide relatively similar and promising results on the image interestingness subtask.

1. INTRODUCTION

Understanding and predicting interestingness of images or video shots have been proposed as a recent problem in computer vision literature [7, 6, 2], which finds many applications such as video summarization [3] or automatic generation of animated gifs [4]. The MediaEval 2016 Predicting Media Interestingness Task is introduced as a new task which consists of two subtasks on image and video levels, respectively. In our work, we concentrate only on the image subtask, which involves identifying interesting keyframes of a given video of a movie trailer, and where we process each frame independently. Details about this subtask including the related dataset and the experimental setting can be found in the overview paper [1].

2. METHODS

Deep convolutional neural networks (CNNs) have revolutionized the computer vision field in recent years, obtaining state-of-the-art results in many different problem domains. In our submission, we tested three different CNN models, which are all based on the popular AlexNet architecture [9]. All of our networks have five convolutional layers and three fully connected layers with a final layer returning a scalar interestingness score. The detailed descriptions of our models are respectively given in Sections 2.1-2.3, and in Section 2.4, we explain how we convert interestingness scores into labels.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

2.1 AlexNet

For our first model, we fine-tune AlexNet [9], which is trained on ILSVRC 2012 task of ImageNet to classify more than a thousand object categories. Image interestingness requires predicting a single real-valued output, so we replace the last soft-max layer with regression layer and use a Euclidean loss layer to fine-tune the model. In our experiments, we only fine-tune the last fully connected layer, while the weights of other layers are not updated. Training lasted approximately 2000 epochs.

2.2 MemNet

Our second model is based on the recently proposed MemNet model [8], which is trained for the image memorability task. Although memorability and interestingness are not exactly the same [5], we think that fine-tuning a model related to an intrinsic property of images could help us to learn better high-level features for the interestingness task. In our experiments, we only update the weight of the fully connected layers, where the training lasted nearly 3000 epochs.

2.3 Triplet Loss

Our third model also follows the AlexNet architecture, but differs from our previous models in that we employ a different training procedure. Specifically, we consider a deep triplet network which is composed of three instances of the AlexNet model where the weights are shared across the instances. We employ a ranking loss similar to that of [4]. However, while the authors of [4] consider a siamese network and a pairwise ranking loss, here we utilize a triplet ranking loss [10] within our network. Once the training is finished, we use a single instance of the feedforward network to predict the interestingness score of a given keyframe.

Considering a triplet network allow us to learn a 1D-embedding space for images, where the triplet ranking loss function enforces an interesting frame to be close by to other interesting frames and far away from the uninteresting ones:

$$L(x, x^+, x^-) = \max(0, D(x, x^+) - D(x, x^-) + M) \quad (1)$$

where x, x^+, x^- denote the anchor, positive and negative samples given as inputs, $D(\cdot)$ represents the distance between the interestingness scores and M represents the margin. In terms of optimization, one critical point is the triplet selection procedure since we observe that using all possible triplets in the training is costly and might lead to a local minima. For this reason, we use a hard negative mining strategy, which is commonly used in similar works. During training, we only fine-tune the fully connected layers

as in our previous models, while all the weights are initialized with AlexNet weights. We interrupted training at the 10000th epoch.

2.4 Interestingness Classification

Thus far, we have described our CNN models which can be used to compute real valued interestingness scores for each key frame of a given video sequence, where these interestingness scores correspond to confidence values. However, the task also requires classifying a frame as interesting or not, in addition to predicting their interestingness scores.

A simple and straightforward way to convert real valued outputs to class labels is to introduce a thresholding procedure. However, choosing a single appropriate threshold value is not easy; in fact, we observe that it is very video sequence dependent. Figure 1 shows the ground truth distributions of the confidence values for the interesting (blue) and uninteresting (orange) frames over all video sequences in the training set. As can be seen, these distributions have a large overlap, demonstrating that a single threshold value won't work.

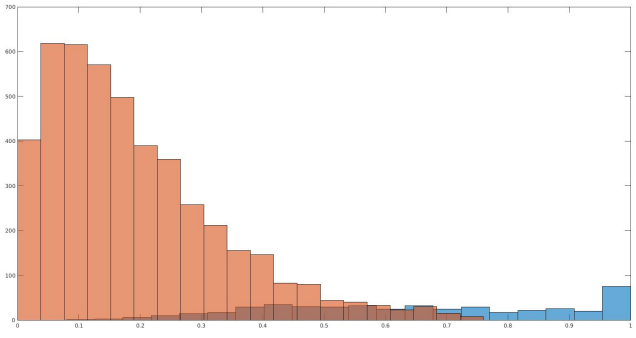

Figure 1: Distributions of the confidence values for interesting/uninteresting frames.

Next, we analyzed the ratio between interesting and uninteresting key frames per each training video. As shown in Table 1, the ratio is, on average, about 1:9. Hence, given a test video sequence, we sort all its key frames according to their predicted interesting scores and classify the top 10% frames as interesting.

Table 1: Statistics for the confidence values for interesting and uninteresting frames over training data

frames	mean	std
interesting	0.11	0.08
uninteresting	0.89	0.08

3. RESULTS AND DISCUSSION

We submit three different runs for the image subtask. While the first run uses our fine-tuned AlexNet, the second one uses predictions from our fine-tuned MemNet model. Lastly, the third run includes the results of our proposed triplet network. All these models are trained by using the provided training data. However, we split it into two as training and validation splits using a ratio of 80% and 20% to deal with overfitting. In our experiments, as the size of the training data is relatively small, we decided to update the weights of only the fully connected layers of the pretrained models.

The performances of our models are evaluated by considering the accuracy and the mean average precision (mAP) scores. Table 2 summarizes our results on the test set. As can be seen, the mAP scores of all the proposed models are not very high, demonstrating that interestingness prediction is not a trivial task. We note that the accuracy values being high are somewhat misleading since the training data is highly unbalanced (see Table 1). Hence, we additionally show the confusion matrices for all of our runs in Table 3.

Table 2: Evaluation results on the test set.

Runs	mAP	accuracy
1	**0.2125**	0.8224
2	0.2121	**0.8275**
3	0.2001	0.8249

Table 3: Confusion matrices for our runs.

Run 1		Run 2		Run 3	
1890	211	1896	205	1893	208
205	**36**	199	**42**	202	**39**

To sum up, the growth of the visual media on the Internet has led to an increased need for understanding and predicting interestingness of images and video shots, and in this work, within the proposed deep models, we treat each key frame of a given video as an independent sample. One possible future direction could be to process each key frame in the context of a local temporal neighborhood or the whole video, by extending our models to process multiple key frames simultaneously. Another extension could be to consider a multi-task learning scheme, which involves jointly classifying key frames as interesting or not and estimating an interestingness score based on a regression-based loss function, which eliminates the need for post-processing the regression scores.

Acknowledgement.
This work is partially supported by the Scientific and Technological Research Council of Turkey (Award #113E497).

4. REFERENCES

[1] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q.K. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingness task. In *Proc. of the MediaEval 2016 Workshop*, 2016.

[2] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. *Proc. International Conference on Computer Vision*, pages 1633–1640, 2013.

[3] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proc. European Conference on Computer Vision*, pages 505–520, 2014.

[4] M. Gygli, Y. Song, and L. Cao. Video2gif: Automatic generation of animated gifs from video. In *Proc. Computer Vision and Pattern Recognition*, 2016.

[5] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1469–1482, 2014.

[6] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *Proc. Association for the Advancement of Artificial Intelligence Conference*, pages 1113–1119, 2013.

[7] H. Katti, K. Y. Bin, T. S. Chua, and M. Kankanhalli. Pre-attentive discrimination of interestingness in images. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 1433–1436, 2008.

[8] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *Proc. International Conference on Computer Vision*, pages 2390–2398, 2015.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[10] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proc. International Conference on Computer Vision*, pages 2794–2802, 2015.