DOI: 10.20396/revpibic2720192770

XXVII Congresso de Iniciação Científica Unicamp
16 a 18 de outubro de 2019 - Campinas | Brasil

What Are Kids Watching at YouTube? Elsagate Detection through Deep Neural Networks

Akari Ishikawa*, Edson Bollis, Sandra Avila

Abstract
Despite YouTube’s efforts to block violent and pornographic content from its platform, it is not prepared yet to deal with the Elsagate phenomenon. As a first to introduce the disturbing cartoons to the literature, we propose along with a dataset a pipeline that classifies Elsagate videos with 92.6% of accuracy and discuss the subjectivity of this problem.

Key words:
Deep Learning, Sensitive Content, Elsagate

Introduction

Although YouTube may seem to be safe for kids, it is today full of Elsagate videos. In the Elsagate phenomenon, we see childhood characters depicted in disturbing circumstances that may be inappropriate for children (e.g., stealing, alcohol, bullying). Although Elsagate channels have existed since 2014, there is no evidence of whether those claims are real, or who are the people responsible for them or their motivation.

Despite the existence of good solutions towards pornography detection, we have few works regarding sensitive content in cartoons and no one focusing on Elsagate [1]. The situation is even direr due to the similitude between Elsagate and non-sensitive cartoons. In this work, we move from the most recent pornography detection literature to propose a Deep Learning-based solution to classify this kind of content in videos.

Results and Discussion

Our first major contribution was the creation of an Elsagate dataset comprised of 285 hours of 1,898 Elsagate and 1,567 non-sensitive videos (Figure 1).

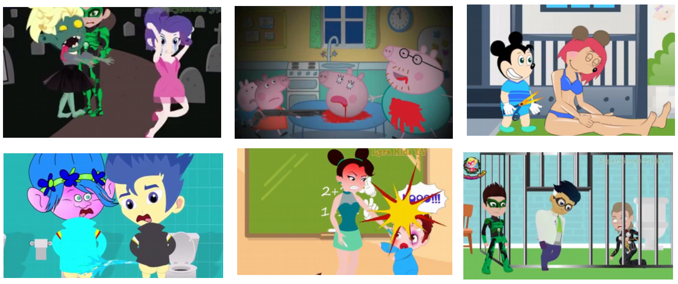

Figure 1: Sample of Elsagate dataset images.

Inspired by the methodology of Perez et al. [2], we present as our second major contribution a complete pipeline to classify Elsagate videos (Figure 2). This approach relies on the video’s frames and MPEG motion vectors for representation, a Deep Learning Architecture (DLA) for feature extraction and a Support Vector Machine (SVM) model for final classification.

Our third but not less important result is the comparison between several Deep Convolutional Neural Networks as feature extractors. We performed experiments with the pipeline above using MobileNetv2, GoogLeNet, NASNet, and SqueezeNet. Our experiments showed that NASNet, when transferred from ImageNet and finetuned to our data, delivered the best results: 92.6% of accuracy and 88.7% of F2-score.

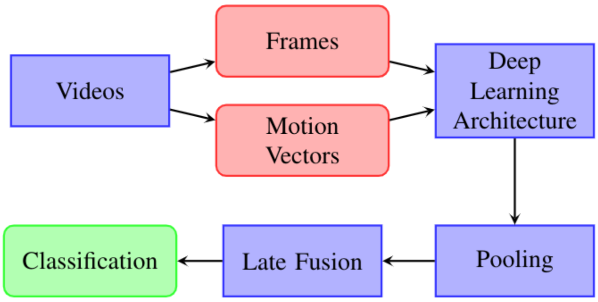

Figure 2: Overview of proposed pipeline.

Despite the good results produced by our solution, we wanted to know if our labels reflected parents’ opinions. Thus, we had volunteers to watch 100 videos from our dataset and answer the question “Would you let a child younger than five watch this video?”. From the 1554 annotations gathered, we had only 70,2% of agreement, showing again how subjective and curious is the problem of stating if a video is disturbing enough to be blocked for children.

Conclusions

In this project, we introduced the Elsagate phenomenon to the literature and proposed the first solution to solve the problem. Despite the satisfactory accuracy achieved, we still find curious how Neural Networks can classify such vague content. Therefore, it is clear the need for a better definition of Elsagate and a discussion about what is suitable or not for children. For future works, we intend to embed the solution in a mobile application and propose a more deep annotation for studying the phenomenon itself.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research and PIBIC/CNPq, FAEPEX (#2555/18, #3125/17), CAPES, FAPESP (2017/16246-0), and Google LARA 2018 for the financial support.

[1] Ishikawa, Akari, Edson Bollis, and Sandra Avila. "Combating the Elsagate phenomenon: Deep learning architectures for disturbing cartoons." arXiv preprint arXiv:1904.08910 (2019).

[2] Perez, Mauricio, et al. "Video pornography detection through deep learning techniques and motion information." Neurocomputing 230 (2017): 279-293.