Violence Detection in Images Using Deep Neural Networks

Edwin-Mark Grigore
University POLITEHNICA of Bucharest
Splaiul Independenței 313, București 060042
edwmrkg@gmail.com

DOI: 10.37789/rochi.2020.1.1.5

ABSTRACT

Software and technology has evolved and expanded so much over the last decades, that is present in everybody's life in every little aspect, and more and more significantly at children's disposal. Starting from this reality, it is necessary the identification of the images that contain scenes of violence or emotionally disturbing scenes, images that contain blood or depict human bodies with open wounds, violent fires, or presence of guns and weapons. Machine learning (ML) is capable of extracting features from images and learn to identify the images that depict inappropriate scenes for children, using different techniques. With the recent advances in deep learning, traditional ML methods, such as Support Vector Machines, have been surpassed by deep neural networks that are also employed by our solution for violence detection.

Author Keywords

Computer vision; violence; violence detection; neural network; deep learning.

ACM Classification Keywords

I.2.10: Vision and Scene Understanding

General Terms

Computer Vision; Deep Learning; Violence Detection.

INTRODUCTION

The common adage "A picture is worth a thousand words" denotes exactly how an image can influence a child, especially if we are talking about inappropriate images. Browne and Hamilton-Giachritsis [1] have shown that aggressive or antisocial behaviour is heightened in children after watching violent television or films. Early exposure to extremely fearful events affects the development of the brain, particularly in those areas involved in emotions and learning [2]. When children see images that are emotionally disturbing, images that depict the world in an inadequate manner for their young minds to comprehend, they can learn fear from situations they should not be exposed to.

In order to prevent the exposure of children to graphic and violent images, these images must be firstly identified. Since parents cannot be physically near their children every single time, nowadays they can rely on the technology they use to achieve this task.

In a World Health Organization report, Krug et al. define violence as "the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, which either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation" [3].

Transposing the notions to the field of images, violence transcends these categories. The term explicit or graphic violence refer to depiction acts of violence in visual media such as film, television, and video games. The violence may be real or simulated.

Graphic violence generally consists of uncensored depiction of various violent acts which includes depiction of murder, assault with a deadly weapon, accidents which result in death or severe injury, and torture.

As the technology advances, Computer Vision leads the way in training artificial intelligence to learn to interpret and understand the visual world. Using deep learning models, we now have machines that can accurately identify and classify objects. Although there is extensive knowledge to develop such deep learning models, only a few have been created that recognize and/or classify the violence depicted in still images, and even less are available for public use.

There are several potential areas that may use machine learning to detect violence, such as parental control applications, and web filtering. Therefore this topic is worth being studied and relevant for computer-human interaction researchers and product designers.

Transfer Learning is used in our work to build a model that detects and classifies violence in still images. This method applies different existing models that have already been trained for general purposes, to the characteristics of the task at hand. First, we must choose from the large pool of the existing deep learning models one to be the basis for our solution. The process of identification of the model that works best on the dataset available is considered to be a key aspect. Second, we take the pre-trained model and use it as a starting point for our violence detection model.

Also, we must decide which layers of the pre-trained model are used in the process and what layers must be built on top of it. Finally, we must adapt and refine the model so it may fit as well as possible the task at hand, process called fine-tuning the model.

The rest of this paper is organized as follows. Section 2 presents the categories of violence that are detected by the

proposed solution. Section 3 introduces the dataset used in training the model. Section 4 presents the approach for violence detection and summarizes the results. Section 5 concludes the paper.

CATEGORIES OF VIOLENCE

Our proposed machine learning solution is intended to identify the following classes. The following lines briefly describe them and how they are connected to violent graphics.

- **Presence of firearms** – the presence of any type of gun or similar fire weapon, whether it is shooting or not, pointed at someone, or threatening a person, regardless of the intent of the subject depicted, is to be classified as violent image.
- **Presence of cold weapons** – any type of melee weapon, ranged weapon or other type of weapon that does not involve fire or combustion is to be classified as violent image.
- **Presence of fire** – any explosion caused by a bomb, any large-scale fire, vegetation fire, any human or animal, living or dead that is burning, any fire caused by a gun is classified as violent.
- **Fight scenes** – any image that represent a fight, regardless of the number of people involved or how they are fighting or the weapons they use, is to be classified as violent image. A fight scene may imply punching, kicking, mutual or from one side. Battle scenes struggles between a person and an animal will also be included in this category.
- **Presence of blood and gory scenes** – any serious body injury, any presence of blood that drains out from a body, any wound or tissue damage, any dead body that shows significant injury, presence of horror creatures, mutant creatures or skull and flesh representation is classified as violent image.

Any other image that is not classified under the above categories will be labelled as **non-violent**.

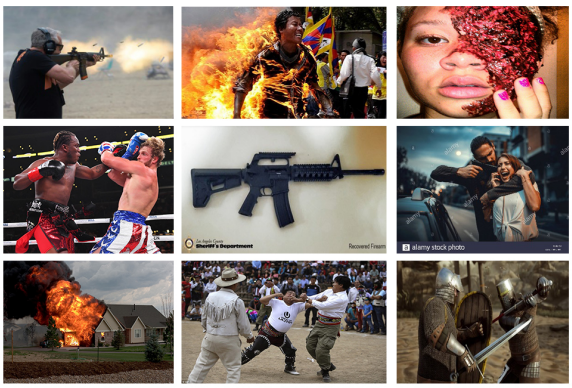

Figure 1. Example of violent images from the dataset

DATASET

A complete dataset is mandatory in order to train the model properly and to achieve good results. Due to the fact that only a handful of violence detection models have been proposed, we were unable to find an existing public database about violence in images with all the categories included. Consequently, we opted for a database of videos to start building our dataset.

Violent Scene Dataset (VSD) created by InterDigital [4-7] is a public dataset for the detection of violent scenes in videos. It is a collection of labels, features, and annotations based on the extraction of violent scenes from films and web videos. It also contains audio annotations of violence-depicting sounds present in the videos.

The dataset consists of 86 short videos downloaded from YouTube and normalised to a frame rate of 25. Also, the dataset contains ground truth created from a collection of 32 films of different genres (which are not included in the dataset due to copyright issues).

The violence identification is made based on two definitions of violent scenes: (1) subjective definition and (2) objective definition. The subjective definition describes a violent scene as a "scene one would not let an 8-year-old child see because they contain physical violence"[1]. The objective definition shows that a violent scene contains "physical violence or accident resulting in human injury or pain".

Frames Extraction

Due to the fact the dataset contains videos, not images, processing work needed to be done. Each frame was extracted from the video, sorted according to the annotation and saved into the new database we created. At the end of the extraction, manual inspection of the resulting set of images was required. Duplicate images and images that are blurry, darkened or where the subject is unclear were removed. Also mislabelled images were moved to the proper category or removed if necessary. In the process of video and image manipulation we used the OpenCV library[2].

The number of images resulted in the process of extraction is in the tens of thousands. However, after a thorough manual inspection and repeated deletion of the unusable files, the database consisted of only around 1000 images, which is rather small for a machine learning solution. Also, different

[1] As presented in the description of the dataset. Available online at https://www.interdigital.com/data_sets/violent-scenes-dataset, last accessed 25 July 2020.

[2] http://opencv.org

perspectives on different categories failed to be gathered into the database, especially for gore and fire presence. In this case, we used Google image search to extend the dataset with graphics that picture the situations that were missing.

Augmentation

Because the dataset is small compared to what a proper dataset would look like, augmentation was helpful to extend the original database. Keras [8] interface allows us to augment the training set after loading the images in memory. It offers multiple ways to do the augmentation, such as image rotation, zooming, cropping, horizontal and vertical flipping, or range shifting.

Another recommended method is mixup [9] used as a regularization technique. Because we do not know the real distribution of data which can lead to overfitting, mixup comes into help to reduce this problem. It introduces combinations of pairs of images and their labels. A shallow explanation is given by this equation, where t is the ratio of mixing two images (a number between 0 and 1):

$$img_{new} = t * img_1 + (1 - t) * img_2$$

SOLUTION AND EXPERIMENTAL RESULTS

In the process of implementation several well-known state-of-the-art technologies have been used. The model was developed and trained using TensorFlow [10].

In Computer Vision, deep learning has been used for tasks such as object identification or scenes recognition. Most solutions employ Convolutional Neural Networks, where lower layers act as feature extractors and the top layers work on the features that are specific to the task. Deep learning models learn different features on their architectural layers. These layers are often connected to a final fully connected layer to get the result. This layered architecture enables us to disconnect the final layer from the network and use the rest of the network as a feature extractor.

An important step in the successful training of the model is choosing the best neural model to apply transfer learning onto. All are state-of-the-art technology, but not all of them suit every problem. It is fundamental to have a model that offers good performance on the task it has been trained on.

In computer vision, several pre-trained neural models have been proposed in recent years. The resulting state-of-the-art deep learning networks are available to the large public and can be used freely and easily, both online or directly integrated in machine learning libraries.

The most popular and best performing such models are: VGG16 [11], InceptionV3 [12], Xception, and ResNet50 [13]. These are the ones we also considered using in developing the model for our task.

On top of these pre-trained models, we built a classifier using the weights from pre-training, consisting of a pooling layer, a few core neural layers and two normalization layers. The final layer is a fully connected layer with 6 (as the number of categories) neurons as output. For the Dense layers we used ReLU activation function, except for the output layer, where we used SoftMax. After splitting the dataset into training set and validation set of 75%-25%, we used the batches generated by Keras and trained the model for various epochs, ranging from 25 to 100. We employed RMSprop with a learning rate of 10-4 as optimizer.

In the process of training the models, the mixup technique helped to deal with overfitting, increasing the validation accuracy by 3-4%. The best performing models were ResNet50 and VGG16. They both provided similar results, but with variable epochs' number (see Table 1). The ResNet model peaks fast, reaching the top validation accuracy after only 13 epochs and maintaining it through the next epochs (up to 100), while VGG16 needs more training time to do so. In general, VGG16 required more time for training with the same batch and dataset size than ResNet50.

Table 1. Accuracy rates on training and validation sets

Epochs / Model	25	50	75	100
ResNet50 without mixup				
Train. Acc.	79.89%	90.56%	91.63%	94.01%
Valid. Acc.	65.34%	71.59%	73.86%	70.85%
VGG16 without mixup				
Train. Acc.	70.31%	79.69%	84.94%	89.23%
Valid. Acc.	73.58%	77.56%	81.25%	80.68%
ResNet50 with mixup				
Train. Acc.	72.54%	82.58%	86.36 %	88.83%
Valid. Acc.	71.88%	65.62%	66.76%	71.02%
VGG16 with mixup				
Train. Acc.	64.11%	73.39%	78.50%	83.52%
Valid. Acc.	69.89%	76.99%	80.97%	80.68%

In Table 2, we can see a comparison between the performances of the models that have been trained with different feature extractors.

Table 2. Performance comparison for different neural model

Pretrained Model	Accuracy
InceptionV3	30%
Xception	41%
ResNet50	74%
VGG16	**81%**

Finally, the model we built using Transfer Learning based on the VGG16 pre-trained model is the one that performed the best, with an accuracy of 81%. Figure 2 shows examples of classification.

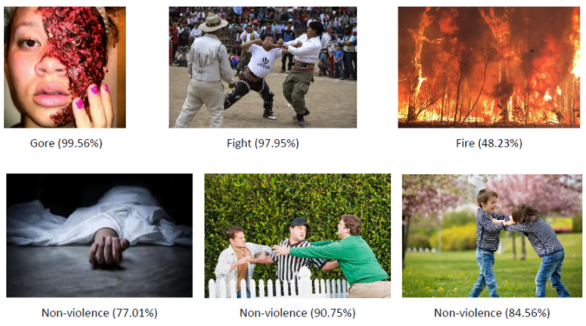

Figure 2. Examples of classification output

CONCLUSIONS

The outcome of this project shows that there is a lot to be done for improving the detection of violent images. Children can be protected using state-of-the-art computer vision technology and, by building a model that would detect the images that can be harmful to see for them and that classify the violence depicted in still images, we believe people can be encouraged to address this issue more. Building machine learning models and using them in all kinds of applications will, eventually, make the world safer for children.

Developing a deep learning model that recognizes violent scenes that would have an emotional impact over an 8-year old child by using deep neural networks is my proposal of work in this field. We built the model by aggregating the knowledge of a pre-trained model and a classification network, with VGG16 being the appropriate state-of-the-art model for the task. The model reports whether a violent or harmful scene is depicted in the image and outputs the class predicted and the score.

Future work will strive to increase the accuracy of the model. This can be acquired by gathering more data and by tuning the model better. Each class has its unique features and there is work to be done to refine the database of each violence category and to identify the features that will increase the accuracy of prediction. As the model will improve, it can be integrated in the parental application that will allow live detection of violence in accessed images.

ACKNOWLEDGEMENT

This work was supervised by Conf. Dr. Ing Traian-Eugen Rebedea, whose advices and guidance helped me in achieving the results.

REFERENCES

1. Browne, Kevin & Hamilton-Giachritsis, Catherine. (2005). *The influence of violent media on children and adolescents: A public-health approach*. Lancet. p. 8
2. National Scientific Council on the Developing Child. (2010). *Persistent Fear and Anxiety Can Affect Young Children's Learning and Development: Working Paper No. 9.* Retrieved from www.developingchild.harvard.edu
3. Krug et al., *World report on violence and health.* Archived 2015-08-22 at the Wayback Machine, World Health Organization, 2002.
4. C.H. Demarty, C. Penet, M. Soleymani, G. Gravier. (2014). *VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation.* In Multimedia Tools and Applications, May 2014.
5. C.H. Demarty, B. Ionescu, Y.G. Jiang, and C. Penet. (2014). *Benchmarking Violent Scenes Detection in movies.* In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), 2014.
6. M. Sjöberg, B. Ionescu, Y.G. Jiang, V.L. Quang, M. Schedl and C.H. Demarty. (2014). *The MediaEval 2014 Affect Task: Violent Scenes Detection*. In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain (2014)
7. C.H. Demarty, C. Penet, G. Gravier and M. Soleymani. (2012). *A benchmarking campaign for the multimodal detection of violent scenes in movies*. In Proceedings of the 12th international conference on Computer Vision – Volume Part III (ECCV'12), Andrea Fusiello, Vittorio Murino, and Rita Cucchiara (Eds), Col. Part III. Springer Verlag, Berlin.
8. Keras | TensorFlow Core. TensorFlow. (2020). Retrieved 2020-06-08, from https://www.tensorflow.org/guide/keras.
9. Zhang, H., Cisse, M., Dauphin, Y., & Lopez-Paz, D. (2017). mixup: Beyond Empirical Risk Minimization. arXiv: abs/1710.09412
10. Metz, Cade (November 9, 2015). "Google Just Open Sourced TensorFlow, Its Artificial Intelligence Engine". Wired. Retrieved 2020-06-08, from https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/.
11. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.
12. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., & Anguelov, D. et al. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2015.7298594
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90