Building a NSFW Classifier

Lucas Ege and Isaac Westlund
Stanford University

Predicting

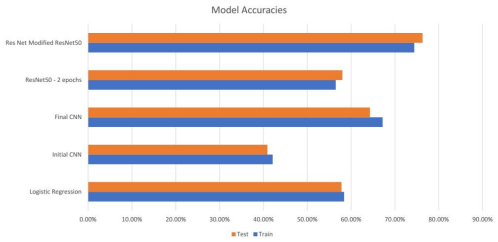
The goal of this project is to develop a classifier to interpret an input image as “safe for work” or “not safe for work” as well as to classify the unsafe images based off of the inappropriate content. Our current categories are gore, pornographic content, and weapons. This is an interesting and relevant application as recently, large social media companies have been working on the same idea as they struggle to curb inappropriate images and videos posted on their platform as they fight to attract advertising.
Our classifier currently achieves 64.3% accuracy on our manually crafted dataset, while industry applications (Google’s SafeSearch API) achieve 77.9% classification accuracy.

Data

Our data has come from two types of sources: Kaggle and manual web scraping. We found decent datasets for knife images and pornographic images on Kaggle, which required minimal preprocessing to download images from URLs and resize. The rest of our data was scraped from the web through direct sites and subReddits through the Imgur API. All our images are colored and normalized to 64x64 with 3 color channels. Utilizing these mediums, we were able to gather ~20,000 images of about 6,000 pornographic images, 4,000 images of weapons, 3,000 gore images, and 10,000 SFW images

Features

Our model uses a Convolutional Neural Network for image classification, taking in a 64x64x3 image as features.
We’ve also implemented an RNN that also takes 64x64x3 images as features, while utilizing recurrences within the network.
Our logistic regression model utilizes the obvious approach of flattening 64x64x3 images into 12,288 long vectors, where each pixel and color value is a feature to the network.



We split our data into 90% training and 10% testing based on our amount of data (20,000 images).

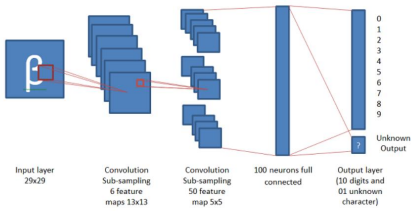


Figure 1. A basic rendition of a CNN [1]

Models

Our initial CNN had two convolutional layers and two max-pooling layers finishing with a fully connected layer with two outputs: SFW or NSFW. We used a sigmoid cross entropy loss function for our loss function simply classifying images as NSFW or SFW. We eventually adapted this model so that the final fully connected layer had 4 outputs that we then fed into a final softmax activation.
We also utilized a larger Recurrent Neural Network using the popular ResNet50 architecture of multiple repetitions of convolutional layers, batch normalization, activation, identity layers and max pooling.

Classifications

NSFW image classification is a subjective qualifier and is subject to change across platforms and over time. We decided that the most important categories of pictures we needed to identify were:

•Pornographic images
•Images showing weapons (knives, guns, etc)
•Images of dead bodies & gore
•Safe images

We decided on this breakdown as it allows the hypothetical end-users of our software to fine tune their platforms depending on what type of content they would allow through. We deemed these classifications the most important in light of contemporary cultures and modern sensitivities.

Future Work

As with most deep learning applications, the quality and quantity of data is vital, so a full scale application would require more data acquisition. We would also like to apply GANs to this field to develop a quality classifier and generator of images together.

Discussion

While our results were not able to match industry giants like Google’s, we were happy to see the level of accuracy we achieved. Data acquisition become a larger task than anticipated, but we were happy with utilizing Imgur as a source because of the wide variety of content sectioned into nice subreddits. In hind sight, we might have been better suited utilizing a pre-built image model and transfer learning to suit our task.
Our results were as good as expected: our data set was able to reach a reasonable size (~20,000 images), where our test set was based off of the same distribution (same sites) of images as our training set, so our model worked well accordingly.

Contact

Lucas Ege
Stanford University
Email: lucasege@stanford.edu

Isaac Westlund
Stanford University
Email: iwestlund@stanford.edu

References

[1]: https://www.codeproject.com/Articles/571462/Multiple-convolution-neural-networks-approach-for