USER-CONTENT IDENTIFICATION OBJECTIVES

EXECUTIVE SUMMARY

In the website hosting and provider business, one of the most difficult challenges in building a successful business according to Convertica, a leading expert in multichannel marketing and business development, is building valuable content. The key to having valuable content is having end-user participation in the content generation process; which not only reduces the amount of overhead for your company by reducing the workload of content generation but also increases the overall volume as well.

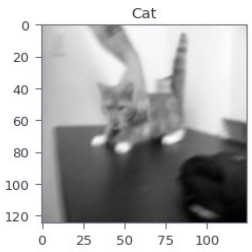
Collecting large volumes of user-generated content has many positives for a start-up website, however, with such large volumes of content, curation becomes a daunting task that is too unwieldy for manual intervention on every piece of new information. To this end, we would ideally have a systematic process in place to automatically identify the user posted content and tag it appropriately.

Having such a system of automatic content recognition for image-based would have enormous benefits. Such as being able to automatically tag and categorize new user-uploaded content so that the site maintains a canonical structure, relevant content can be automatically placed under relevant sub-sections on the site, and the information flow is not bottlenecked by a manual process that keeps the site moving organically around the clock. An additional benefit of such a system would include flagging inappropriate materials automatically so that we don't accidentally alienate a large portion of our user base by a single bad actor uploading inappropriate content.

RESEARCH DESIGN

For this problem, we used a sample user-generated dataset used for a machine learning competition on a well-known data science hub, Kaggle.com. This dataset contains an image database of 25,000 high-resolution images of cats and dogs in various settings, backgrounds, and angles. The images in this set are pre-labeled with the correct description in the file, so we can use this information to validate our model against it.

For this research, we are going to optimize our model for accuracy agnostic of training time. We will run models with different parameters, as well as with various image resolutions. The baseline version of this model will scale the images from various resolutions to static 75 by 75 images, and then to 125 by 125 images to test the accuracy effects of lossy image compression/scaling.



We will also control for various aspects of the neural network, including the convolutional network layer sizes, pooling layers, optimization and loss functions.

TECHNICAL OVERVIEW

The model construction and benchmarking methods were conducted purely in the cloud, leveraging a pre-canned environment from a world-class provider of machine learning solutions, Google, called Colabratory. This cloud environment enables us to execute research more efficiently, while also allowing maximum reproducibility by taking out the variability of an individual desktops hardware configuration for the benchmark results. We can also publish our research globally and allow any astute reader to reproduce our results for themselves in a sandboxed environment.

Additionally, for this research, we also leveraged an industry-leading machine learning framework, TensorFlow, which Google also produces. TensorFlow gives us access to the same underlying technology that powers several of the most advanced analytical systems in production today.

Using this framework, we set up an artificial neural network algorithm for training our classification system. For this problem, we chose a Convolutional Neural Network (CNN), algorithm that is trained on the cats and dogs dataset with various parameters so that we can conduct research on the optimal training configuration, in which we reach maximum accuracy, agnostic of the time spent on the algorithms I/O and computational heavy training phase.

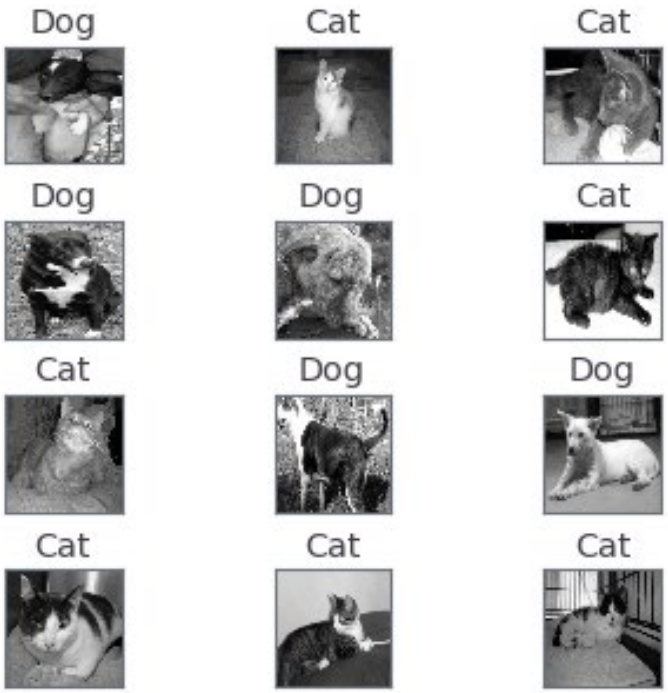
At the heart of this study is the aforementioned Convolutional Neural Network algorithm, and three tuning parameters: learning rate, training epochs, and batch size. We will execute the "baseline" algorithm with the minimum specifications to get the algorithm to recognize the images to some degree of confidence. We will use a standard split, train and test procedure which uses 20,000, or 80% of our 25,000 images, to train the model and the remaining 20%, or 5,000 images, to validate our model.

There is an independent test set supplied along with this data that we can use for additional visual inspection of how well the model classifies the images based upon our visual inspection. The modeling procedure also randomly shuffles the dataset during the preparation phase, this way each time we run the process we can freshly train all of the models on the same data as we compare our results.

CONCLUSION

With the base case scenario, we used images scaled to 75x75 pixels and two separate instances of the CNN model trained on the same set of data. The first instance trained in 9.89 minutes and the second version trained in 10.82 minutes, the main difference between the two versions being that the second one has two extra convolutional layers.

The baseline version of the model passes quick visual inspection, however, upon benchmarking the model against the 5,000 test images we ear-marked at the beginning of the research, we see that it only correctly identifies 2,490 of the 5,000 images; which is not even as good as a coin-flip. The second version of the model with additional convolutional layers performs significantly better, accurately predicting 3,769 of the test images, or 75.38%.



These results are satisfactory, however, accurately predicting 3 out of every four images is not the results we were hoping for. After the first batch of results, we rest the model to scale the images down to 125 by 125, a 60% increase in data capacity.

Training times scale approximately linear with the images size, as the baseline model trains in 28 minutes, and 31 minutes respectively.

The results of our modeling accuracy improved to a respectable 49.22% in the baseline and 78.28% in the robust second version. Therefore, given the thirst for accuracy, I recommend we move forward with implementation of the more robust model trained using higher-resolution images as the data source.

For additional information on this research, and how it is conducted, please visit the Colabratory Notebook.