Binary Classification of Beautiful vs. Non-Beautiful Photographs Using a Deep Convolutional Neural Network

Stephen Crawford, Class of 2022

The purpose of this research was to use a deep convolutional neural network (DCNN) to create a classifier which would then be analysed for insight into the differences between beautiful photographs and non-beautiful photographs. This insight is important for the answering of the philosophical question "what is beauty?", the valuable guidance it offers in how artists may create beautiful works, and as further research into a popular line of inquiry.

In order to train a model to differentiate between beautiful and non-beautiful photographs, it was necessary to have ample samples of both classes. In order to create as representative dataset as possible, photos which had won recognition in any of six prestigious photography contests were used as stand-ins for beautiful photos. The six contests used are the Deutsche Börse Photography Foundation Prize; the World Press Photo Contest; the Hasselblad Award; the Sony World Photography Awards; the National Geographic Photo Contest; and the Fine Art Photography Awards. In order to build the non-award-winning portion of the dataset, a simpler approach was adopted. Specifically, the first 200,000 photos of the LabelMe database were downloaded before a random subsample of 12,000 images was selected. This was a quick and efficient method of building the non-award-winning portion of the database.

For reasons beyond the scope of this summary, a convolutional neural network was used to extract features from the samples. These features were then fed into a densely connected network. This design was determined the best option for binarily classifying photographs.

For this research, the model was to be considered effective if it could differentiate between winning and non-winning photos at a rate higher than 0.5 or 50%. Given the even dataset, the model could be expected to guess the label of a sample correctly 50% of the time without knowing anything about the sample. With respect to this naive baseline, all version of the model performed well.

In order to verify the functionality of the model, a confusion matrix was manually calculated. For each iteration of the model, the precision, recall, and F-measure were calculated by tallying the results of the model's predictions over the 6000-sample test set. Interestingly, it appears that the model struggles to classify the same images that a person may have difficulty classifying. In a given dataset subsample, the classifications it mislabels are often those which are somewhat ambiguous or won awards from the Fine Art Photography Contest where more abstract submissions are allowed (*Figure1*).

In order to validate the model, the same architecture was trained on an entirely new dataset. While the main dataset included photos across many categories, the second dataset included only architecture photos. This smaller dataset included 900 award-winning photos and 900 non-award-winning photos. Significantly, the model trained on the smaller but more specific dataset outperformed the model trained on the more diverse dataset. All statistics of the model trained on only architecture photos show better performance than the model trained on the entire set of award-winning photographs. For further information please contact: Stephen Crawford (scrawfor@bowdoin.edu).

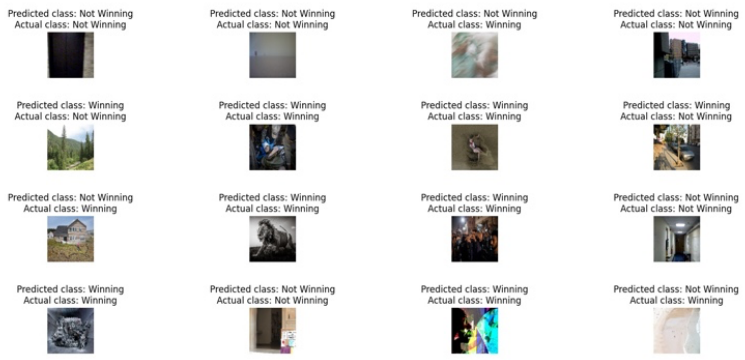

Figure 1: Predictions made by the model for a sample of photographs.

Funded by: **Gibbons Research Fellowship**