VANGOGHORNO: Identification of Van Gogh's Paintings

Ron Fan (ronbo), Belinda Li (lib49), Nathan Wong (wongnat)

October 17, 2017

VANGOGHORNO is a dataset containing paintings from European and Russian artists ranging from the 13th to 20th centuries, including hundreds of paintings from Dutch Post-Impressionist painter Vincent van Gogh.

The problem we're tackling is categorizing paintings as by van Gogh or not by van Gogh; specifically, deciding based on visuals alone whether a painting was likely to have been created by van Gogh. This is an interesting problem that, for humans, is not trivially easy, but can be done with fairly high accuracy by experts with sufficient knowledge and information - which we supply to our machine learning algorithm, in this case. Moreover, it is interesting as there are many useful applications of this problem, including painting authentication for auctions and art collection, helping art historians study the unique style and characteristics of van Gogh's art, allowing laymen to quickly get some knowledge on an unfamiliar painting, and being able to classify paintings of unknown origin via artist.

We believe that, in terms of difficulty, this problem is appropriate for machine learning because it presents a binary classification where the positive label is clearly separate from the negative label (i.e. it was drawn by a specific person). However, it's also not too easy, because despite being drawn by the same person, the paintings produced by van Gogh reflect changing styles and influences over the duration of his life.

To ensure the dataset is high quality, we took a few steps. First, we ensured we had at least a few hundred images in each category (van Gogh/not van Gogh). Second, we made sure that our non-van Gogh dataset was as comprehensive as possible by including paintings from every century from 13th to 20th. We included paintings similar in style to van Gogh's, such as those from other impressionist painters of the 19th century, as well as paintings very different in style to van Gogh's, such as those by late medieval (13th century) artists. We can also know that our data and classifications are accurate as the images have been verified by experts at the van Gogh museum. Finally, to ensure consistency, we resized our images to approximately the same size.

All of our input data are JPEG images, and the data in them can be read by loading the image in a buffer and reading in the pixels one by one.

To divide our data into training and tests, we randomly selected 20% of all our images and labeled them as test data. Randomness helps eliminate possible biases in splitting the data, biases that may affect how well the the test data represents the training data. We want the test data to be as similar as possible to the training data (without being exactly identical) in order to evaluate whether learning has occurred.

We believe our dataset is ethical because it consists of artworks in the public domain that had digital images released online with the full intention of being viewed by art admirers around the world. The artworks contain no sensitive information, and any machine learning work done on the data set is highly unlikely to cause offense or harm towards others due to the uncontroversial nature of the material. We are unable to envision any use of an algorithm for suggesting whether a painting was created by Vincent van Gogh that could be construed as unethical.