HACKRICE

Title: Data Science

Date: 09/22/2017

Author: Abhijeet Mulgund

Monet or MoNot?

Link to join competition!

Track description:

With paintings selling for hundreds of millions of dollars, telling real paintings from forgeries is a high-stakes business. This year at HackRice, you'll have the option to work on a computer vision task, and be judged on your model's ability to tell whether two paintings were painted by the same artist. In order to perform well in this task you'll need to leverage recent advances in the field of computer vision as well as your own ingenuity and creativity. From the movement of brushstrokes to the use of light and dark, successful algorithms will likely incorporate many aspects of a painter's unique style. For more information see the original Kaggle competition page.

Track prompt:

You have 12,747 different paintings from the Impressionist and Post-Impressionist eras along with artist info for each. Your task is build a model or algorithm to determine if any two paintings are by the same artist. Your model will be scored on paintings not included in the training data.

Keep in mind that the solution must be written by you. While you can use other sources for help, you **cannot** just copy-paste any code other than whatever was provided with this track. You also **cannot** include the test set in your code in any direct or indirect way. **Any solution violating these two rules will be immediately disqualified.**

Track Evaluation:

Your solution will be scored on its binary logloss on a private held-out test set. The test-set is split in a 30-70 fashion, the former of which is used for scoring on the public leaderboard and the latter of which is used for calculating the final score.

Kaggle In-Class:

We'll be using Kaggle In-Class as the platform for this competition. You'll use this page to download the data, submit your predictions for scoring, and discuss with other competitors.

External Data & Sources:

You are **not allowed to use any raw external data** to help with this competition even if it's publicly accessible. Raw external data includes (but is not limited to):

1. Data Sources other than the provided dataset

2. Labels for competition data not provided by HackRice

Using external data and/or sources indirectly is permitted given that you post it in the External Sources and Pre-trained Models Kaggle discussion thread. Indirect External Data and Sources includes (but is not limited to):

3. Pre-trained models
4. Specific algorithms designed by others (Specific enough to **not** have its own Wikipedia page)

When in doubt, just post it in the thread. **Any solution found to violate any of these rules will be immediately disqualified.**

Ideas for how to get started:

Before coming up with a solution you should first explore the data. Try different things like

1. Examining the frequency of certain artists, genres, etc. in the dataset
2. Visualizing a sample of the training data

Once you have a feel for the data, go ahead and try out some machine learning algorithms. We recommend using siamese neural networks. You might start with naively training your algorithm, but to do well, you will most likely have to be creative with how you train your model.

Sample code/startup solution:

Go to this kernel on Kaggle to see a basic example of loading the data and training a siamese neural network.

References:

1. The Original Kaggle competition page
2. WikiArt.org
3. What are Siamese neural networks, what applications are they good for, and why?
4. A neural algorithm of artistic style
5. How Do We See Art: An Eye-Tracker Study