Title: Car Classification by Invariant Risk Minimization
Category: General Machine Learning
Team members: Abby Audet (aaudet/006173095) and Annette Jing (ajing/06351705)

Motivation: What problem are you tackling? Is this an application or a theoretical result?
We will by applying machine learning to identify the make, model, etc. of a car based on an image for the purposes of detecting fraud in online car sales and automating aspects of car maintenance and repair. Successful implementation of this project could open the door for more advanced uses in the future, such as having a traffic camera identify cars matching descriptions of local Amber Alerts.

Method: What machine learning techniques are you planning to apply or improve upon?
A problem with car images is that certain makes have existed a lot longer than the others, so an naive ML algorithm might learn to not classify newer looking models to these makes.
When encountered with pictures of new models from these makes, this can lead to gross misclassifications.
Similarly, when we pool images from several different sources, spurious correlations like that described above can arise due to unforeseen factors such as a database containing more images taken from a certain angle.
We will use the Invariant Risk Minimization (IRM) to prioritize learning invariant features across data collected from different "environments."
The environments here are decided by the decade the car model is from and which source the sub-dataset it belongs to is taken from.
By implementing invariant risk minimization, we hope to improve the accuracy of our algorithm by teaching it to not only ignore but also identify potentially confounding factors.

Intended Experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?
We plan to apply IRM with various loss functions and compare it with standard empirical risk minimization based ML algorithms.
For example, we may train and validate the models with images of cars made before 2000 collected from several databases and test their performance with images of cars that are made after 2000 collected from a different database.
Relevant Datasets:
- The car connection picture dataset
- Vehicle make and model recognition dataset
- Stanford cars dataset
- Vehicle image database

Prior Research:
- Invariant Risk Minimization (Arjovsky et al.)
- Causal inference using invariant prediction: identification and confidence intervals (Peters et al.)