The Summer Undergraduate Research Fellowship (SURF) Symposium
2 August 2018
Purdue University, West Lafayette, Indiana, USA

Investigating Dataset Distinctiveness

Andrew W. Ulmer, Kent W. Gauen, Daniel P. Merrick, Zohar R. Kapach,
Yung-Hsiang Lu
Department of Electrical and Computer Engineering, Purdue University

ABSTRACT

Just as a human might struggle to interpret another human's handwriting, a computer vision program might fail when asked to perform one task in two different domains. To be more specific, visualize a self-driving car as a human driver who had only ever driven on clear, sunny days, during daylight hours. This driver – the self-driving car – would inevitably face a significant challenge when asked to drive when it is violently raining or foggy during the night, putting the safety of its passengers in danger. An extensive understanding of the data we use to teach computer vision models – such as those that will be driving our cars in the years to come – is absolutely necessary as these sorts of complex systems find their way into everyday human life. This study works to develop a comprehensive meaning of the style of a dataset, or the quantitative difference between cursive lettering and print lettering, with respect to the image data used in the field of computer vision. We accomplished this by asking a machine learning model to predict which commonly used dataset a particular image belongs to, based on detailed features of the images. If the model performed well when classifying an image based on which dataset it belongs to, that dataset was considered distinct. We then developed a linear relationship between this distinctiveness metric and a model's ability to learn from one dataset and test on another, so as to have a better understanding of how a computer vision system will perform in a given context, before it is trained.

KEYWORDS

Deep Learning, Computer Vision, Dataset Bias