International Journal of Multimedia Data Engineering and Management
Volume 13 • Issue 1

Metamorphic Testing of Image Classification and Consistency Analysis Using Clustering

Hemanth Gudaparthi, University of Cincinnati, USA*
Prudhviraj Naidu, University of Cincinnati, USA
Nan Niu, University of Cincinnati, USA

ABSTRACT

Testing deep learning systems requires expensive labeled data. In recent years, researchers began to leverage metamorphic testing to address this issue. However, metamorphic relations on image data remain poorly understood. To gain a deeper understanding of these metamorphic relations, the authors survey common image operations modeling covariate shift, manually classify and categorize the underlying metamorphic relations, and conduct experiments to validate the classifications. In these experiments, the authors trained three popular convolutional neural network architectures on an image classification task. Next, the authors applied metamorphic operations on input test images and measure the change in classification accuracy and cross-entropy loss. A hierarchical clustering algorithm cluster these results and plots a dendrogram. The authors then compare the groups from manual classification and the clusters from the algorithm to provide key insights. The authors find that Affine and Noise relations are consistent. Furthermore, the authors recommend metamorphic relationships to save time and better test deep learning systems in the future.

KEYWORDS

Clustering Analysis, Metamorphic Testing, Neural Networks

INTRODUCTION

Deep learning (DL) has proliferated not just in research papers but in many walks of life. For example, our ongoing work exploits DL in predicting combined sewer overflows Gudaparthi et al. (2020), Challa et al. (2020) and Matlibe et al. (2021). If data is the new oil Bhageshpur (2019), then DL systems have become the *de facto* refineries. One such area where DL has shone the brightest is computer vision. The culprit for this rise are convolutional neural networks (CNNs). In 2012, Ciresan et al. (2012) used CNNs to achieve near-human performance of image classification on the MNIST dataset. From self-driving cars Bojarski et al. (2016), through detecting objects from aerial images

DOI: 10.4018/IJMDEM.304390

*Corresponding Author

Xia et al. (2018), to photo beauty mobile applications Xu et al. (2019), the CNN-based computer vision solutions have become widespread. As with all software systems, DL systems come with their own bugs and fallacies.

As one of the premises of DL is to raise the level of intelligence thereby requiring fewer human validation of the system, the bugs can lead to greater disasters. For instance, in a 2018 Tesla Autopilot crash, the National Transportation Safety Board found the crash occurred due to "limitations of the Tesla Autopilot vision system's processing software to accurately maintain the appropriate lane of travel" National Transportation Safety Board (2020). In another case occurred in China, a facial recognition system tagged a woman as a jaywalker, while she was never actually there at the intersection Liao (2018).

These cases illustrate why thoroughly testing DL systems has become critical. However, software testing often requires labeled data, which is a costly resource commonly expended in training DL systems rather than testing them. Moreover, data labeling is time-consuming and can sometimes be error-prone. Additionally, real world data can change before DL system can be tested.

To address the bottleneck of lacking sufficient labeled data, researchers began to leverage *metamorphic testing* Chen et al. (1998), a property-based software testing technique useful for alleviating the oracle problem Lin et al. (2018) and for generating new test cases. The prototypical example of metamorphic testing is the program that computes the sine function Segura et al. (2016): The exact value of sine (x) could depend on how floating-point computations are handled in the specific implementation, representing an instance of the oracle problem. Metamorphic testing uses properties like sine (x) = sine (π−x) to test any implementation without having to know the concrete values of either sine calculation, i.e., without knowing the test oracle of sine (x) or the test oracle of sine (π−x).

Properties like sine (x) = sine (π−x) are known as metamorphic relations (MRs) Lin et al. (2021). Each MR consists of two parts: (1) an input transformation that can be used to generate new test cases from existing test data, and (2) an output relation that compares the outputs produced by a pair of test cases Segura et al. (2016). As far as image data is concerned, Ding et al. (2016) developed five MRs based on input biological-cell images to validate the open-source light scattering simulation software that performs discrete dipole approximation: altering the image size, shape, orientation, refractive index, etc.

Despite the emergence of metamorphic testing on image data, little is known about the relations of the MRs applied to DL testing. For example, the MRs in DeepXplore Pei et al. (2019) focus on the decision boundaries of multiple CNNs, whereas DeepTest Tian et al. (2019) and DeepRoad Zhang et. at (2018) depend on change in single neural network's output. To shorten the gap, we manually classify several MRs on the image data, and then perform automated clustering analysis based on the DL performances of the MRs. In particular, we train three CNN architectures (MobileNet V2, NasNet Mobile, ResNet50 V2) by using the Keras library Chollet et al. (2015) and by exploiting a subset of the ImageNet data. By comparing the manual classification and MR clustering results, we offer concrete insights into the relations and choices of metamorphic testing in the context of CNN-based image classification.

This paper makes three main contributions: we classify 11 MRs into five groups based on the underlying image data operations, we embed the MRs into the three CNN implementations on a subset of the ImageNet data, and we perform clustering analysis to uncover the distance and natural groupings of MRs. Our work can guide machine learning developers' choosing representative MRs to test DL solutions, ensuring the functional and nonfunctional qualities. In what follows, we present background information and related work in Section II. We then manually clarify the MRs for the task of image classification in Section III. Section IV details our CNN implementations, Section V describes the clustering analysis, and finally, Section VI concludes the paper.