Background.

This project builds upon on-going research to model and simulate galaxy collisions within the Computational Science Ph.D. Program and Physics and Astronomy Departments here at Middle Tennessee State University.

Galaxy Zoo: Mergers is a citizen scientist effort to find models for 62 target galaxy collisions with the use of SPAM. SPAM is an O(N) simulation that models galaxy collisions and recreates the tidal distortions or shape that the galaxies undergo during collision. Citizen scientists were presented with target images of two colliding galaxies and several simulations that attempt to model the target collision. Using the innate human ability to visually recognize patterns, citizen scientists identified sixty-six thousand models that have resemblance to the target image. Once selected, these galactic models also underwent a tournament-like scoring system to identify the best models for a particular target system, and give them corresponding human fitness scores for how well each individual model matched the target image.

Project.

The human ability to process visual patterns and identify similarities is incredibly complex. Trying to identify and measure the “similarity” between two images is often arbitrary and could depend on hundreds of different factors depending on the objects or images being compared. Even with the increase of computer image processing and computational power, previous efforts to manually program and predict the similarity between two galaxy images has had poor results with the ongoing galaxy collisions research. We hope to utilize neural networks to capture the human ability to predict similarities, at least in the narrowed application of galaxies.

The focus of our project is to create a variety of methods for predicting how well a model “fits” a target image using neural networks. Neural network regression models will be trained on pre-existing model images and associated human fitness scores from Galaxy Zoo: Mergers. The regression model will take galaxy images and produce a machine fitness score that attempts to predict human machine scores. The performance of our networks will be based on how well the final machine scores statistically correlate with human scores for a set galactic model.

The primary goal is to develop a neural network regression model for each individual target system. Each target system can be trained on hundreds to thousands of model images and human scores. In addition, a variety of neural network architectures can be implemented and analyzed. Perhaps a particular architecture performs the best across the majority of target systems. To evaluate the performance of our neural networks, the primary metric will be the statistical correlation between neural network predictions and human fitness scores. In addition to the correlation, graphs can be constructed to identify outliers, or possible overfitting to training data. Once the regression model is built and operational, new model images could be generated and humans can agree and disagree with the predicted score and improve the process.

Another possible application is to build a general-purpose neural network model that can predict the similarity between any two pictures of galaxies in collision. This could be trained on a sampling of models and target images from all of the target systems above. This may be difficult to build, as each system undoubtedly has variance in how their similarity scores scale. For example, a 0.75 similarity for one system may resemble a much higher “similarity” between images than a 0.75 for another system which contains poorer quality models and data. However, the flexibility may be well worth the effort. A general-purpose comparison could be applied to future galaxy collision systems.

Another application could be to create single-layer networks for all of the target systems. With these single-layer networks, the weights could be visualized and scientists could identify what sections of the image are most crucial for predicting a similarity score. For example, do all the systems have a greater importance along tidal features or the edges of the galaxy shape? Is pixel importance related to the distance between the axis between galaxy centers? Which sections of the images can be mostly ignored?

One ambitious idea is to incorporate existing open-source image classification models. WNDCHRM is an open-source image classification software, that goes beyond the raw pixel intensities. Instead WNDCHRM extracts hundreds of image features, identifies which image features best predicts an images classification, and uses those metrics to train and predict a classification model. Perhaps WNDCHRM could be utilized for its automated feature extraction and identification, then use those image feature to build an even more accurate and robust neural network regression model.