Predicting PSA Grades for Baseball Cards

A generalization for CNN Architectures in Sports Analytics

Group 8

Hyun Suk Ryoo
School of Data Science
University of Virginia
Charlottesville, VA 22903
hr2ee@virginia.edu

Anoop Nath
School of Data Science
University of Virginia
Charlottesville, VA 22903
nux9aq@virginia.edu

Cepehr Alizadeh
School of Data Science
University of Virginia
Charlottesville, VA 22903
ca3eh@virginia.edu

Marin Lolic
School of Data Science
University of Virginia
Charlottesville, VA 22903
ejz2sg@virginia.edu

December 7, 2022

Abstract

Sports trading cards are a substantial market, and the evaluation of a card's condition is currently done by human experts. While the evaluators are professionals, the existing process is subjective and time-consuming. We take a deep learning approach to this problem, training several neural networks to predict the PSA grades of baseball cards based on photographs. Our most promising model is ResNet50, which achieves a 92.6% validation set accuracy after hyperparameter tuning. This model could be used to supplement human judgment, possibly reducing the time and monetary cost of evaluating the condition of baseball cards.

1 Motivation

The sports trading cards market has experienced a large boom in recent years and is currently estimated to be a 1.3 billion-dollar industry. The price of a single trading card can go into the millions of dollars, with the most expensive card recently being sold for $12.6 million. There are several factors that determine the value of a sports trading card. These include:

- Player and team: The more popular the player and/or team, the more value the trading card has.
- Rarity
- Condition
- Graded vs non-graded: There are a handful of companies that will judge the authenticity and condition of a card. A higher graded card will be worth more.

The most popular card grading service is Professional Sports Authenticator (PSA). They review submitted cards for authenticity and then grade the card on a 1-10 scale based solely on the condition of the card. Cards that are in mint condition will receive higher scores, whereas cards with damage will receive a lower grade.

The process of getting a card graded is currently not simple. It requires one to send the trading card to a third-party company via mail. The card is then inspected manually by a human and given a numerical grade (usually on a 10-point scale) and then sealed in a card holder and send back to the owner. The whole process takes several weeks and is an expensive endeavor.

In our project, we created a deep learning model that can predict the PSA grade of a card, on a 1 to 10 scale, using just a single image of the front of the card. A successful model has several benefits for both professional grading services and sports trading card enthusiasts.

The model can be used as an additional or alternative grading method for professional grading services. This could help them improve accuracy, reduce grading time, decrease costs, and decrease human errors and biases.

An additional benefit is for collectors and traders to get a grade estimate, even if unofficial, for an ungraded card. For example, a collector considering purchasing an ungraded card could be more confident about the value of the card if they can get a quick estimate for the grade of the card.

2 Dataset

The dataset that was selected to address the PSA Grades of Baseball Cards can be found here. The dataset is a collection of images sized at 150x200. To accomplish this similarity in images, original images were rotated, scaled down, or padded using replication of pixel values. In our dataset there are 10 classes. In the dataset repository a folder structure is given where each subdirectory of the root directory is named psa1 through psa10. The folder structure follows the rule where the folder psaX contains all the images for psa grade X in the folder psaX.

The dataset was collected through ParseHub and Collectors.com. ParseHub is a free interactive web scrapping tool where users can interact with the tool to select which portion of the page to scrape data from. The source that was used to collect the data by ParseHub was collectors.com. The companies mission statement is as follows. "Helping collectors pursue their passion through industry leading grading and authentication, tools to help collectors research and find the next big thing, and marketplaces to buy and sell." The grading system found on this platform can be trusted as industry leading and reliable, which will be our source of truth for our project. Although, some can state that PSA grading can have mixed opinions, PSA grading is a thorough and costly effort, which require professional PSA graders that are selected and trained to maintain quality of each card.

For a summary of the dataset that will be utilized in this study, there are 1,150 cards for each grade. Therefore, our entire dataset will be a collection of 11,500 images evenly distributed for the grades of 1 through 10

3 Literature Review

3.1 Quality Detection and Control

Computer-based statistical methods have been used in the field of industrial quality control for at least the past decade. Examples include Li et al. (2014), who utilized principal components analysis to find defects in mobile phone cover glass. One highly-cited paper is Wang et al. (2018), who trained a deep convolutional neural network with the goal of detecting defects in a wide variety of production processes through images. Their network bears resemblance to AlexNet (2012), and uses a total of nine layers, including convolutional layers, pooling layers, and two dense layers. Their network makes use of a variety of training and regularization techniques, including downsampling images, ReLU activation functions, dropout and L2 regularization. The network achieved a 99.8 percent test accuracy rate on the DAGM defect dataset, which was state-of-the-art at the time of publication.

Yang et al. (2019) take a more complex approach to the problem of detecting faulty welding joints, training a generative adversarial network to create additional images due to the lack of existing data. They then use a Yolo-V3 deep convolutional neural network with 24 layers to detect and identify various problems in welded metals. One part of the network identifies whether a problem exists, while a second part of the network locates the problem on the picture. Finally, Ashok et al. (2014) trained a probabilistic neural network (PNN) to classify fruit by quality, achieving a test set accuracy of 88 percent for defective versus non-defective prediction. While all the previous research we found performed binary classification (faulty vs. non-faulty), it should be simple to adapt this into a multiple classification problem using a softmax output layer.

3.2 Deep Learning and Sports Analytics

Data science in sports has existed in some fashion or another for decades. Sports such as baseball have embraced it since Bill James' ingenuity in the 1980s and its application was popularized through Michael Lewis' novel *Moneyball: The Art of Winning an Unfair Game*, which centered around using data to identify lesser known and inexpensive talents in the sport. Although James' sabermetrics inspired analytics that are widespread in some of today's sports, the application of this science to continuous sports such as soccer was difficult. However, advancements in deep learning to enable it to handle high dimensional data and the newfound efficiencies to gather data like high-resolution video have allowed for a new age of sports analytics (Tuyls et al., 2021). By combining computer vision with statistical learning and game theory, a new frontier is suggested in the *Game Plan: What AI can do for Football, and What Football can do for AI* in issue 71 of the Journal of Artificial Intelligence Research (JAIR). The culmination of these three aspects in science is a suggested Automated Video Assistant Coach (AVAC) system that can provide predictive analytics for players, coaches, and fans.

Computer vision enabled by deep learning approaches plays a pivotal role in the obtainment of the data for the AVAC system. Human pose estimation can be leveraged to determine a player's graphical representation geographically while object detection can help with the tracking of players and the ball. Systems like AVAC would learn a player's skill vectors and improve the predictive modeling associated with the player, their team, and the game itself. Limitations, however, do exist with improving the performance of computer vision in this realm. A challenge exists with the view of the field from the broadcast cameras being focused on where the ball is, which can leave the majority of players out of view. Some stadiums are outfitted with hundreds of cameras that can capture most imaginable angles of the players of the game but this is by no means a standard. This challenge can be overcome even in the most basic stadiums by adjusting the optical properties of the cameras. Further, accompanying the video with audio and text modalities creates more information content to be used by downstream applications.

A specific model put forth by this paper was its model characterizing the patterns of penalty kickers in the English Premier League in the 2016 - 2019 seasons. An 18-dimensional vector is created for each player analyzed that aims to describe their on-field attributes. Clusters of players are then detected using K-means clustering. The dimension of the 'Player Vector' is then reduced from 18 to 3 using PCA and finally the Jensen-Shannon divergence is measured between the Nash probabilities of the clusters created. This model sits at the lowest level of the soccer analytics suggested by the journal: goal-scoring optimization. This is, however, intrinsically and inherently related to the other levels of analysis which are match and championship level respectively.

4 Method

With the given literature review in the sport analytics domain and the quality prediction in images, our main intention is to see if currently existing models will generalize well to our dataset of baseball grades to predict their PSA grade. In many classification problems there are many innate differences between two types of images. For example, in a picture of a cat or dog, there are differences that easily differentiate a cat from a dog such as physical features. However, in our dataset all cards are standardized. All cards have the same boundary and similar appearances. There are experts, PSA graders, that are professionally trained to find these differences and grade each card.

Because our data is images, convolutional neural networks are the obvious choice of machine learning method. Our experiments include generalizing CNN models such as Le-Net, AlexNet, and ResNet to see if the models can find the differences in grading. Our methodology includes testing the architectures of the CNN models to see how accurately they can classify cards into PSA grades.

Another aspect that we considered is the dataset itself. We were initially concerned about possible differences in the size or structure of the images, and we considered the possibility of resizing images or using pixel replication. However, the images were sufficiently similar that we did not need to further process them before model training.

4.1 ResNet

The ResNet50 architecture was chosen to be tested due to its solution to the vanishing gradient problem by relying on shortcut connections. These connections create identity mappings that allowed deeper models to exist without the risk of higher training error. Coupling this advantage with ResNet50's reliance on batch normalization made an attempt at leveraging this architecture worthwhile.

4.2 AlexNet

Unlike the ResNet50 implementation, AlexNet is a much shallower architecture that consists of only 8 layers. While we wanted to attempt using a deeper CNN architecture, we thought it was also important to test our results using AlexNet as well. The architecture's reliance on data augmentation was a feature we wanted to experiment with. Similar to ResNet50, AlexNet's tendency to produce shorter training times were a characteristic we deemed important.

4.3 Le-Net

Working backwards chronologically, the group finally tested one of the original CNNs with the Le-Net architecture. Le-Net will also be the shallowest architecture we implement as it only has 5 layers. Although its simplicity can be advantageous for explainability, it is also a shortcoming as it may have insufficient capacity for this task.

5 Experiments

Before testing any of the chosen architectures, we first split the dataset of 11,500 images into a training set of 9,200 images and 2,300 images for a validation set.

The first architecture that was tested was the 50-layer ResNet50 implementation using a ReLu activation function, the Adam optimizer leveraging a learning rate of 0.001, and was trained on only 100 epochs. To prevent overfitting, a dropout layer with 0.5 frequency was added as well as an average pooling layer. This first attempt at tuning the architecture produced results that were very impressive. A 89.00% accuracy on the validation set was achieved with this preliminary run. Reviewing the results of each epoch, it's shown that the accuracy didn't drastically improve after roughly somewhere between th 15th and 20th epoch which is also an interesting observation for our team as we originally even believed 100 epochs to be too little. As expected based on the team's research into the architecture, the training time of the model was quite long. With the results of the ResNet50 preliminary experiments exceeding expectations for accuracy.

The AlexNet model was tested next and with an additional 19 layers that were added. These layers are a mix of convolutional layers, dropout layers, simple dense layers, and batch normalization layers. The architecture was experimented with as the unaltered AlexNet is rather shallow and additional layers was thought to aid its performance.

Lastly, the LeNet model was tested next with an architecture that had more layers than our ResNet50 model but not as many as the AlexNet implementation. Two convolutional layers with ReLu activation were added in addition to two pooling layers. Two dense layers with ReLU activation and a dense layer with a softmax activation were included. This deeper architecture was tested in hopes of bringing a higher accuracy but it did not bring much success compared to the ResNet50 architecture tested.

Hypertuning for all the models were conducted as well. All models under went a grid search with the combination of the two hyperparameters. The learning rate was investigated for values of [0.1, 0.01, 0.001, 0.0001] and the optimizers that were investigated were Adam and Stochastic Gradient Descent. Upon doing some hyper parameter turning of all the models we were able to generate the accuracy measures of all models as shown in table 1. The hyperparameter is denoted as a combination of Learning Rate & Optimizer.

6 Results

Hyperparameter Tuning Metrics			
Hyperparameter	ResNet50	LeNet	AlexNet
0.1 & Adam	0.095217	0.095217	0.097391
0.1 & SGD	**0.926087**	0.101304	0.099130
0.01 & Adam	0.638261	0.186087	0.101304
0.01 & SGD	0.910870	0.095217	0.652174
0.001 & Adam	0.890000	0.279565	0.646957
0,001 & SGD	0.669565	0.281304	0.503913
0.0001 & Adam	0.886087	**0.413913**	**0.681739**
0.0001 & SGD	0.424783	0.365652	0.099130

Table 1: Table 1: Hyperparameter tuning of models

From table 1, we were able to see that for ResNet50, the best performing hyperparameter was a learning rate of 0.1 and optimizer of SGD, which yielded an accuracy of 92.61%. For LeNet the best performing hyperparamter was a learning rate of 0.0001 and the Adam Optimizer, which yielded an accuracy of 41.39%. For AlexNet the best performing hyperparameter was a learning rate of 0.001 and the the Adam optimizer, which yielded an accuracy of 68.17%

7 Conclusion

Our results showed that a deep learning model could accurately predict the PSA grades of baseball card using just a single image of the front of a card. After training three different deep learning structures using a dataset consisting of 11,500 images of baseball cards, the best performing model was ResNet50, which had a 92.6% accuracy on the validation set. This model significantly outperformed the two other models that were tested: LeNet and AlexNet.

Even though the results are promising, the current accuracy is not high enough to fully replace human grading in official baseball card grading. Further improvements to the model would be needed before it can be useful for official PSA grading. An accuracy of 98% or higher could incentivize them to incorporate deep learning models to their official grading process, either as a primary or secondary source to complement human grading. One way to make this improvement would be to have various images of the card. Our dataset only had a single image of the front of the card. Having numerous images, including the back of the card, could easily boost the accuracy.

However, our model could still be used by baseball card collectors to get a quick estimate of the PSA grade of their cards. The current alternative is to send their card to get officially graded, which currently costs anywhere from $50 to over $500 and has an estimated turnaround time of up to 90 days.

An extension of this project could be to train a deep learning model to not only calculate a PSA grade but also to estimate the dollar value of the card. As mentioned in the introduction, the PSA grade is only one factor in determining how much a card is worth. A model could be trained to identify a card (player, year, company of the card manufacturer etc.) as well as the PSA grade and then compare it to similar cards that have recently sold in the market. Such a tool would be beneficial for baseball card collectors and traders who want to get a quick estimate for how much a card is worth. The same framework for our work could easily be extended to other types of cards such as football, basketball, and even non-sports cards such as Pokémon and Yu-Gi-Oh.

8 Member Contribution

Max focused on the experiments section. He was able to construct the structure for ResNet, AlexNet, and Le-Net to do some experiments and ingest the data into the models. He was also able to do a grid search on the hyperparamters to find the best performing model.

Anoop focused on the motivation of the project as well as the conclusion derived from our results and their implications.

Cepehr produced the literature review of the works related to Deep Learning and Sports Analytics. He also contributed to the "Method" and "Experiments" sections.

Marin focused on the literature review, particularly the section on Quality Detection and Control. He also wrote the abstract and contributed to the "Method" section.

References

[1] Tuyls, K., Omidshafiei, S., Muller, P., Wang, Z., Connor, J., Hennes, D., Graham, I., Spearman, W., Waskett, T., Steel, D., Luc, P., Recasens, A., Galashov, A., Thornton, G., Elie, R., Sprechmann, P., Moreno, P., Cao, K., Garnelo, M., … Hassabis, D. (2021). Game plan: What ai can do for football, and what football can do for ai. Journal of Artificial Intelligence Research, 71, 41–88. https://doi.org/10.1613/jair.1.12505

[2] Li D, Liang LQ, Zhang WJ. Defect inspection and extraction of the mobile phone cover glass based on the principal components analysis. *Int J Adv Manuf Technol.* 2014; 73(9-12): 1605–1614.

[3] Wang T, Chen Y, Qiao M, Snoussi H. A fast and robust convolutional neural network-based defect detection model in product quality control. *Int J Adv Manuf Technol,* 2018; 94: 3465-3471.

[4] Yang L, Liu Y, Peng J. An automatic detection and identification method of welded joints based on deep neural network. *IEEE Access 7,* 2019: 194592 - 164961.

[5] Ashok V, Vinod DS. Automatic quality evaluation of fruits using probabilistic neural network approach. *International Conference on Contemporary Computing and Informatics (IC3I),* 2014: 308-311.