CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Rehaan Ahmad; Brian Yang

Project Number

S1501

Project Title

Machine Learning Based Citrus Orchard Health Analysis with Autonomous Drone Technology

Abstract

Objectives/Goals

Our goal is to build an algorithm that can detect an orange disease with an accuracy of more than 90% and have a processing time of less than 10 seconds per tree. The solution should present data in a graphical and intuitive manner to the farmer. The drone should also be able to map out a whole orchard autonomously. Our algorithm (written in python using TensorFlow) should begin by location oranges on trees, and then classifying these oranges as normal or deformed.

Methods/Materials

We used a DJI Phantom 3 drone, and developed our own application to autonomously control it. We also used a laptop to process the algorithm. Our algorithm consists of two parts, locating oranges, and classifying the oranges as normal or deformed. To locate oranges, we developed a custom OpenCV (image processing library) based color algorithm to locate oranges on a tree and feed that information to our machine learning algorithm. We also created a custom convolutional neural network trained with 10,000 oranges (5,000 normal oranges and 5,000 rotten oranges) to classify each orange. To present the data, we developed a Tkinter (Python UI library) based interface to display a heatmap of the orchard to help visualize its health. Each tree is represented by a square grid where the color of the square is indicative of the ratio of deformed oranges.

Results

After training our network, we tested the classification accuracy of our model with 2000 images. Our algorithm achieved 98.8 percent accuracy. We also tested our network over different training data sizes of 2000, 4000, 6000, 8000, and 10000 oranges. After plotting the accuracies, there is a trend of increasing accuracy with a bigger dataset. This process shows that with an even larger dataset of 20,000 oranges, the accuracy can reach above 99%. We also tested classification time with sets of 200 test oranges. Our algorithm was able to process 200 oranges in 12 seconds. In our test cases, each orange tree had around 50 visible oranges, which took three seconds for our network to process.

Conclusions/Discussion

Our solution provides comprehensive data to the farmer in a cost-effective and accurate manner. This method helps identify outbreaks such as citrus greening, citrus canker, or any disease that affects the fruit with 98.8 percent accuracy. This method can also be trained to analyze the health of other fruit orchards such as apple and lime.

Summary Statement

We created an algorithm that can count the number of diseased and normal oranges on each orchard tree, and created a drone application to map out a whole orchard.

Help Received

We wrote all of the code and designed the entire algorithm ourself. However, we got in contact with Professor Won Suk Lee from the University of Florida to get an idea of existing technologies.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Akhil Arun; Sahana Srinivasan

Project Number
S1502

Project Title
A Machine Learning Based Approach to Skin Lesion Segmentation Using Superpixels

Abstract

Objectives/Goals
Lesion segmentation plays an important role in the early identification and treatment of melanoma, a deadly skin cancer. Our objective was to develop an end-to-end methodology for segmenting skin lesions that was independent of dermoscopic image resolution and size in order to allow for increased efficiency and greater flexibility in implementation.

Methods/Materials
We employed the SLIC pre-processing algorithm to split images into smaller, spatially coherent areas called superpixels, and we extracted two feature sets that both included texture properties and distance metrics, with one set also containing RGB values and the other a color histogram. Random forest classifiers, dense neural networks, and two distinct convolutional neural networks were designed and tested to find their optimal configurations, after which we employed a post-processing, hole-filler algorithm to improve segmentation predictions.

Results
Our color histogram-based random forest achieved the highest accuracy of 89.83%, and our color histogram-based dense neural network obtained an accuracy of 89.01% but represents a more efficient methodology. Both of these superpixel-based segmentation techniques demonstrated comparable accuracy levels to prior studies done on a pixel level while processing 145 to 4,070 fewer pixels, making the process significantly more efficient and allowing it to operate on a wide array of images.

Conclusions/Discussion
We propose a pipeline that includes superpixels and our color histogram-based feature set, a random forest for machines with sufficient computing power or our dense neural network for those without, and the implementation of our post-processing algorithm to achieve efficient and high-accuracy segmentation of dermoscopic images regardless of resolution. Our results indicate that segmenting dermoscopic images using a superpixel-based approach can perform comparably to machine learning techniques on a pixel basis, even at the faster speed and that superpixels can potentially be used in a wide variety of medical image analyses to increase efficiency and flexibility while maintaining accuracy.

Summary Statement
We developed a computational model for segmenting skin lesions that is independent of dermoscopic image resolution and size, achieving a 90% accuracy comparable to that of previous studies while increasing efficiency and flexibility.

Help Received
Our fathers, software engineers, helped us set up our PC and helped in the initial selection of the programming language.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Mythili Bhethanabotla

Project Number

S1503

Project Title

Automated Classification of Rhabdomyosarcoma Tumors: A Novel Method to Determine Appropriate Treatment

Abstract

Objectives/Goals

The survival rate of children with Rhabdomyosarcoma (RMS) cancer is under 30%. Among other challenges, existing research does not accurately classify between the subtypes of this disease. In fact, sarcomas may be misclassified in up to 25% to 40% of cases, which results in harmful under treatment and a lower chance of survival. Accurate classification of the subtypes of this disease will allow doctors to assign patients to the appropriate procedures and even save lives. Therefore, I propose a novel methodology to segment and accurately classify the two histopathological subtypes (Embryonal RMS and Alveolar RMS) of Rhabdomyosarcoma cancer.

Methods/Materials

I developed a computational pipeline that consists of three complementary procedures: segmentation, registration, and classification using a deep learning framework. Using a database of 1764 MRI and DWI scans after data augmentation, I segmented the tumor from each scan. I worked on a novel segmentation technique based on the level set method that can be used to segment 3D images. Once the scans were segmented, the MRI and DWI for each patient were registered using a complex registration algorithm. Registration is the process of fusing two images together in order to form a more comprehensive representation of the image. These registered images were the input to the deep learning framework. I used a Convolutional Neural Network (CNN) in order to classify the images as Embryonal RMS versus Alveolar RMS. After running the images through the network I created, I obtained an accuracy rate for the classification of the images.

Results

I visualized the weights of the first and second layer of the CNN, and the weights had smooth filters without any noisy patterns, indicating that the network is converged and well-trained. The results also show an 85% accuracy for the network using the pretrained weights and a 97% accuracy for the network trained from scratch.

Conclusions/Discussion

I present the first study to implement the fine-tuning of a pretrained CNN model on multimodal brain tumor image datasets. This methodology can be used by radiologists in practice when classifying a the type of cancer and deciding which treatment to give to the patient in order to increase their chances of survival against RMS. This method can be used for classification of other cancers and has the potential to replace manual classification, thus saving time and resources.

Summary Statement

I devised a novel methodology using segmentation, registration, and deep learning techniques in order to accurately classify between the histopathological subtypes of Rhabdomyosarcoma cancer.

Help Received

This project was conducted at the Stanford University's Department of Radiology under the supervision and guidance of Dr. Imon Banerjee and Dr. Daniel Rubin.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Cynthia Chen

Project Number

S1504

Project Title

Fighting Malnutrition: Automated Optimization of Nutrition Using a Novel n-Dimensional Linear Programming Algorithm

Abstract

Objectives/Goals

Malnutrition is prevalent issue in this world: it accounts for 58 percent of all mortality. As a result, my project aims to solve the problem of malnutrition using the method of linear programming (LP) optimization. Mathematically, the problem of nutrition can be modeled using a n-dimensional integer linear program (ILP). My main objective for my project was to develop and create a novel ILP algorithm which returns an accurate and optimal solution while minimizing cost.

Methods/Materials

Previously, other mathematicians have developed algorithms such as the Simplex or Interior Point algorithms to solve a LP problem. However, I found that a big disadvantage of using those algorithms is that there exists the case where the algorithm yields infeasibility and a null solution, but an optimal solution exists and is not found. Consequently, I focused on increasing the feasibility of the LP by developing a novel algorithm which is able to resolve the no solution (infeasibility) case mentioned above while still maintaining the accuracy of the returned optimal cost-effective solution. After analysis, I found that previous algorithms were unable to relax the constraints when they are too restrictive. When developing a new algorithm to address this problem, I used a novel method of weighted constraints that I implemented myself. I created an n-level deep recursive algorithm that was able to successfully decrease the probability of infeasibility decreases, and return an accurate optimal solution. Also, I was able to integrate my model with Wolfram Mathematica in order to receive form input for user info (age/gender), as well as use its built-in ILP function.

Results

In this project, I was able to successfully create an improved model with a novel algorithm which is able to reduce to probability of infeasibility and optimize the n-dimensional linear program more accurately. Overall, the model is able to successfully optimize nutrition while minimizing cost in order to help solve malnutrition.

Conclusions/Discussion

The model that I developed can serve as a primitive model for other ILP problems, such as transportation networks, production planning optimization, or any problem which involves a cost-benefit analysis. In the future, I would like to generalize the model and algorithm I developed in this project, and integrate the model into a mobile software application.

Summary Statement

In my project, I developed a novel n-level deep recursive linear programming algorithm in order to solve the problem of nutrition.

Help Received

I completed most of the work myself. My math teacher, Mr. Bradley Stoll, helped me with using Wolfram Mathematica. Also, Dr. Gary Blickenstaff was the official sponsor for my project.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Hagop J. Chinchinian

Project Number

S1505

Project Title

Algorithmic Radio Data Classification and Pulsar Detection Using MATLAB

Abstract

Objectives/Goals

The objective is to automate the scoring process for Fast Fourier Transform plots, which represent a radio signal as observed by a telescope. The algorithm calculates a score for the subplots which constitute an FFT plot (the method by which humans manually analyze FFT plots) and classifies the source of the signal as a pulsar, radio frequency interference (RFI), and noise.

Methods/Materials

MATLAB was used in order to develop the algorithm (along with the Statistical Analysis Toolbox and the Image Processing Toolbox).

The development of the algorithm focused on identifying patterns in the Fast Fourier Transform plots, assigning scores based on these patterns, and classifying the plots into a radio source (pulsar, RFI, or noise).

Results

The algorithm was tested with 200 FFT plots, and the algorithm had an 81.8% similarity to the human rating. A 95% confidence interval was calculated, showing that the true percent similarity between the algorithm and human rating is between 77% and 88%.

A chi-squared test of homogeneity was applied to determine if the radio source classifications were accurate (testing if the distribution of algorithm and human classifications were the same). The chi-squared test showed no significant difference between the distributions of algorithm and human classifications.

Conclusions/Discussion

Testing the algorithm showed that it was relatively accurate in scoring the FFT plots and determining the signal source.

In order to improve the algorithm and its accuracy, it is suggested that a database be established for which the algorithm can refer to when scoring plots. This will enhance the algorithm when determining the scores/classifications of plots that do not distinctly show patterns representing pulsars, RFI, and noise. Machine learning algorithms may also be implemented.

Summary Statement

An algorithm was developed in order to analyze Fast Fourier Transform plots, which represent radio signal data, by automatically scoring the plots on a scale (used when manually scoring such plots) and determining the source of the signal.

Help Received

The algorithm was fully designed by myself, and Mr. Paul Mekhedjian answered my questions regarding the mathematics behind the algorithm. The data was obtained from the Pulsar Search Collaboratory (PSC). Dr. Raymond Ellyin provided general feedback on the project and is the PSC adviser at my school.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Vennela Chukka; Anusha Fatehpuria

Project Number

S1506

Project Title

A Deep Learning Approach for Gland Segmentation in Cancerous Tissue Images

Abstract

Objectives/Goals

The goal of this project was to use a deep learning based fully convolutional neural network (FCN) for pixel level gland segmentation of cancer images, and test the ability of the model learnt on colon cancer images to segment glands in different types of cancer images.

Methods/Materials

The gland FCN model is initialized by transferring the VGG-16 learnt model, pre-trained on an image database of 1.5 million images. Matlab training and testing scripts in the FCN-Matconvnet toolbox were modified for gland segmentation, and the training was done on a computer with a GPU for 1200 epochs (iterations) over 40 hours of training.

Results

The algorithm was trained and validated on a colon cancer image database, for which the training and testing accuracy were approximately 96% and 90% respectively. The same model was also tested on a prostate cancer image dataset, for which the segmentation results qualitatively look accurate.

Conclusions/Discussion

Overall, the results from this model are promising, especially since this accuracy was achieved with a training set of less than 100 images, proving that it is possible to construct a highly accurate deep neural network model to automatically segment a multitude of glands in various cancer type tissue images.

Summary Statement

This project aims to develop a deep learning algorithm based on fully convolutional neural networks and transfer learning to segment glands in cancer images, then test its ability to be able to expand to different types of cancers.

Help Received

We contacted various pathologists for cancerous tissues images. After downloading the FCN-Matconvnet scripts, we contacted professionals to get more information about the specifics regarding the code. After writing the code, we requested a professional to simply run the code.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Siddharth Das

Project Number
S1507

Project Title
A Fast Efficient Technique for Finding a Path through Multiple Destinations

Abstract

Objectives/Goals
In a real world situation, members of different organizations often need to travel to multiple destinations without passing through the same place twice. However, identifying an efficient path to travel through multiple destinations can be extremely difficult and time-consuming problem. In this work, I propose an algorithm/technique to find an efficient path covering many destinations, within a short amount of time. The path starts from the source location and goes through all the destination locations. This heuristic uses Dijkstra's algorithm to find different paths between different pairs of destinations and then determines the Least-Cost path for each segment. This approach allows people and organizations to save time, fuel and money. As a result, this technique saves environment as well.

Methods/Materials
I have completed this project in Windows laptop that has Chrome Browser. Javascript programming language and Dracula graph library were used to implement the core path-finding algorithm. I used html language to develop the web-based application for accepting user#s input locations and few other settings. Google Maps and Directions APIs were used to display the final output in the web-browser.

Results
I have run several sample test situations that involves between 5 and 50 destination locations. My data shows that the algorithm produces an efficient path within 31 to 372 seconds, depending on the number of destinations. For many such test situations, I have visually verified that the resulting paths are the best possible paths that a human could have generated by using trial and error over several hours of time. These data and observations confirm that the proposed algorithm is efficient and fast.

Conclusions/Discussion
I developed an algorithm/technique to quickly find an efficient path that covers multiple destinations. For many test situations, I have visually verified that the resulting paths are the best possible paths. For the problem involving 50 locations, this technique takes only 372 seconds to compute the path (exhaustive approach for 50 locations would have taken years). I conclude that the proposed approach is ready for real-world usage in the organizations and companies. Since the user requires very little time and knowledge, this approach will be quite appealing and useful to many users in the community.

Summary Statement
A novel technique to efficiently find a path covering multiple destination locations, within a very short amount of time.

Help Received
I designed and programmed the algorithm myself after studying different relevant algorithms.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Rishi Desai

Project Number

S1508

Project Title

New Link Grouping Network Visualization Technique for Social Network Analysis and Biological Network Alignment

Abstract

Objectives/Goals

Traditional node-link diagrams for visualizing networks do not scale as networks become larger and are nicknamed "hairballs". To address the need for the visualization of large networks, the Institute for Systems Biology, Seattle, developed BioFabric, a software tool that depicts nodes as horizontal lines and edges as vertical lines. However, the BioFabric default layout does not utilize link relations. My objective is to develop link grouping network visualization that allows the user of BioFabric to closely analyze link relations, as well as network connectivity and biological network alignments.

Methods/Materials

I implemented the link grouping network visualization within the existing open-source BioFabric Java code base. I developed the link grouping algorithm that places edges of the same link relation adjacent to each other forming link groups. I tested the link grouping layout on a Facebook social network and the alignment of the protein-protein interaction (PPI) networks of mouse and plant.

Results

The link grouping layout displays each social circle in the Facebook social network in organized fashion that amplifies similarities and differences in connectivity. Social circles with high connectivity are easily distinguishable due to the similar shape of the edges laid out on each node-line. The link grouping layout displays the network alignment of the mouse and plant PPI networks with its aligned edges and unaligned edges placed in separate link groups. Based on the relative width of each link group, one can see how topologically similar these networks are, thus allowing the transfer of biological data between the mouse and plant PPI networks.

Conclusions/Discussion

I developed a novel method to visualize large and complex networks that provides several advantages over traditional node-link diagrams and the BioFabric default layout. This method gives researchers a powerful tool to analyze link relations, network connectivity, and network alignments through simple inspection.

Summary Statement

I developed a software tool based on link grouping that allows researchers to visualize and analyze large and complex networks through simple inspection.

Help Received

Mr. Longabaugh at Institute for Systems Biology, Seattle, and Prof. Hayes at UC Irvine provided guidance and valuable comments.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Jason Z. Dong

Project Number

S1509

Project Title

A Novel Algorithm to Diagnose Medical Issues: Identifying Facial Paralysis via Convolutional Neural Networks

Abstract

Objectives/Goals

To (1) compress the existing VGG-16 Convolutional Neural Network into a smaller neural network specifically tailored to a reduced number of categories of images to decrease training times and (2) to accurately identify the presence of facial paralysis in new images.

Methods/Materials

First, open-source images were collected containing faces with and without facial paralysis. Images consisted of faces spanning a variety of ages, in different lightings and angles. These images had different degrees of facial paralysis present, ranging from slight signs of drooping to easily distinguishable deviations from normal faces. After collecting these images, JPEG files were then converted into an LMDB database using a portable MacBook. These databases were then processed through the convolutional neural network, where unsupervised learning was able to take place. Final results with this algorithm were attained using a desktop computer with an Intel Core i7-6700K Processor.

Results

After training this network on only a small set of images, results consistently achieved higher than a 95% accuracy rate with rates often exceeding 98%. These results were able to be achieved in small amounts of time, not exceeding 100 milliseconds with a reduced training time of under four hours.

Conclusions/Discussion

Focused on faces with and without facial paralysis, my algorithm was able to effectively determine whether or not new faces had any signs of facial drooping while also minimizing the training time necessary to learn key features in trained images. As impending strokes only provide a few key features, with facial paralysis being one of these identifiable signs, the immediate identification of facial drooping is imperative to avoid the neuronal death associated with strokes. Satisfying the two main objectives of this project, the algorithm I developed provides an effective way to classify images and diagnose medical issues, particularly facial paralysis for impending strokes, in a timely manner with low training times. This artificial neural network is capable of learning and identifying any type of image, being able to attain an accuracy rate consistently higher than 95%. Able to effectively classify images, this algorithm can accurately diagnose other medical issues by iteratively learning through known symptoms.

Summary Statement

I developed a deep convolutional neural network able to efficiently and precisely diagnose medical issues such as the presence of facial paralysis to identify impending strokes.

Help Received

Mr. Jason Lee helped to facilitate the application process of this project and conducted meetings through regular check-ups.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Aratrika Ghatak

Project Number

S1510

Project Title

Developing a Predictive Model for On-Campus Crime Using Machine Learning Algorithms and Reporting via Mobile App

Abstract

Objectives/Goals

On-Campus crime is a leading issue in US Colleges, a serious concern for parents and a difficult challenge for the college authorities. US Department of Education captures data related to various types of crimes in US colleges. In my project, I wanted to leverage that dataset, collect other related data from various government agencies and then run several Machine Learning algorithms to find the best predictive model. Finally, I will empower prospective students and authorities with a mobile application to easily access the trend and prediction.

Methods/Materials

The project has been conducted in 3 phases. First, I gathered on-campus crime data at each individual crime type level and related demographic, social and economic census data from various websites. I delegated each college a Crime Severity Index(CSI), which is a weighted score considering severity of each type of crime. Depending on the CSI of each college, I assigned them a safety grade from A+ to F. The second phase was to run various Machine Learning algorithms with my training dataset, compare the key metrics and to come up with a best fit. I ran Decision Tree, Bayes Net and Logistic Regression on my training data set which is approximately 1.5M data points. The last phase was to develop a mobile app for users to access the historic crime statistics at detailed level, trend and predicted results. I used WEKA as Machine Learning Software, MIT AppInventor2 as app development platform and Google Fusion Database as cloud platform to store information.

Results

Based on the key measures of various algorithms, Decision Tree was found to be the best fit and could predict on-campus safety grade correctly for 87%. It also had better accuracy, precision and recall values. This validation was done using 10-fold cross-validation technique. Moreover, it was important to remove the outliers so that the model does not over-fit.

Conclusions/Discussion

In conclusion, It is evident that law enforcing agencies, college security and college authority can take great advantage, using machine learning algorithms like Decision Tree to effectively fight crime on the campus and residence hall. My project also educates the prospective students, parents and other visitors in the college by providing a handy mobile application to check the detailed crime statistics, its influencing factors, historical trends and future prediction.

Summary Statement

My project demonstrates that Machine Learning can effectivity be used to create a predictive model for campus safety grade and provides an important tool to fight against on-campus crime.

Help Received

I have collected necessary data, consolidated and trained the model by myself. Thanks to my teacher, Mr. Higgins for supporting and sponsoring my project and thanks to The University of Waikato, New Zealand for making WEKA freely available for everybody to use.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Dhruba J. Ghosh

Project Number

S1511

Project Title

Artificial General Intelligence: Efficient Reinforcement Learning to Achieve Human-Like Reasoning

Abstract

Objectives/Goals

Artificial neural networks suffer from a phenomenon known as "catastrophic forgetting", which prevents artificial intelligence agents from learning multiple tasks. This project explores various network structures and training methods to deal with this barrier to achieving general intelligence.

Methods/Materials

Used Python programming language with Theano and Keras libraries. Various unconventional artificial neural network structures were assessed on tasks such as object recognition and cursor navigation based on training speed and resilience.

Results

The traditional sequential models demonstrated catastrophic forgetting to a large extent, achieving over 90% accuracy on one task but under 10% on the other. Meanwhile, a sequential model trained simultaneously on both tasks reached only 50-60% accuracy. However, progressive or 2-column neural networks improved about 20% on this benchmark. There was a significant increase in performance and training efficiency.

Conclusions/Discussion

Progressive neural networks are an efficient way to improve AI performance on multiple tasks. They can be easily extended to encompass more than two tasks, through the use of multiple columns, or subnetworks. Choosing the output (since PNNs create multiple outputs) is a trivial task that may be left to a separate machine learning model. With learning models such as the PNN, the issue of catastrophic forgetting is no longer a major obstacle in achieving general intelligence.

Summary Statement

I explored various training techniques and artificial neural network models to improve model resilience when given more than one task to learn.

Help Received

None. I designed and wrote all the programs myself. I used the Keras and Theano scientific computing libraries in Python.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Hannah Hu

Project Number
S1512

Project Title
Image-Based Air Quality Prediction Algorithm

Abstract

Objectives/Goals
Air pollution is a concerning global issue with increasing public awareness and serious health consequences. As air quality and smog levels fluctuate throughout a city within a matter of hours, with pollutants circulated by wind and trapped by buildings, it is difficult to know whether it is safe to breathe the air. Air quality index (AQI) values representing the concentration of particles as small as 2.5 micrometers are recorded by sensors. We attempt to estimate the AQI using only the visual characteristics of an image representing the region.

Methods/Materials
The single image haze removal with dark channel prior algorithm uses the assumption of the #dark channel# in the three color channels within a window to estimate the levels of atmospheric degradation per color channel, which is then used to remove the haze and create a depth map relative to the amount of atmospheric interference calculated. This calculated atmospheric value can be used to estimate the AQI. Using a set of images covering years# worth of smog in Beijing, China, alongside data gathered from the Beijing US Embassy covering hourly AQI values, we created a model of the relationship between the atmospheric values outputted by the dehazing algorithm from the images and the corresponding AQI values.

Results
Our results show that there is relationship between the atmospheric and AQI data, but easily loses accuracy in its predictions when any source of error is introduced through discrepancies in the data, such as different locations or times. With adjustments made to flag known sources of error, an accurate linear model was developed to predict AQI levels based on the characteristics of the given image.

Conclusions/Discussion
The ability to quantify smog levels from an image introduces many possibilities with which this data and process can be used or improved upon for the benefit of smog-afflicted citizens.

Summary Statement
I developed a targeted algorithm to use only the visual characteristics given by a digital image representing a region to quantify and estimate the smog levels as the air quality index of the region.

Help Received
I was advised and mentored by Dr. George Cutter from the South West Fisheries Science Center, a branch under the National Oceanic and Atmospheric Association.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Amy Jin

Project Number

S1513

Project Title

Deep Learning-Based Automated Tool Detection and Analysis of Surgical Videos to Assess Operative Skill

Abstract

Objectives/Goals

Annually, 7 million patients suffer surgical complications, many of which are linked to inadequate surgical training and a lack of individualized feedback on how to improve operative technique. To improve surgical quality, it is essential to assess operative skill, a manual process that is time consuming, subjective, and requires experts. Real time automated surgical video analysis would provide a way to objectively and efficiently assess surgical performance. Thus, we aimed to create a deep learning model to perform surgical tool detection, which would enable us to automatically track and analyze tool movements, giving rich insights into operative skill.

Methods/Materials

Since tool localization had not previously been done before, we first had to assemble our own dataset for this task. We designed a MATLAB annotation interface and used it to label 2200 video frames from 15 cholecystectomy surgical videos with the coordinates of spatial bounding boxes around tools. This dataset was then used to train a deep learning model to perform automated tool detection and localization. We leveraged region-based convolutional neural networks (R-CNNs). Input video frames were passed through the VGG-16 convolutional neural network (CNN) and then through a region proposal network (RPN), and the model outputted the spatial coordinates of bounding boxes around surgical instruments.

Results

In comparison with state-of-the-art approaches for automated tool detection, we outperformed existing methods by 23%, improving mean average precision (mAP) from 63.7 to 78.2 using just 10% the amount of training data. Additionally, the network's processing speed at deployment is 5 fps, achieving real time surgical tool detection. We further demonstrated the ability of our method to assess surgical quality by extracting and analyzing key metrics that reflect surgical skill level. In particular, we examined tool usage patterns, tool usage times, movement range, economy of motion, and path length, and used these measures to evaluate surgical performance.

Conclusions/Discussion

To the best of our knowledge, this study was the first to develop an approach for automated surgical video analysis. We collected a new dataset with the spatial bounds of tools, leveraged region-based convolutional neural networks for real time surgical tool detection and localization, and automatically extracted key metrics to assess operative skill.

Summary Statement

We developed a unique approach for automated surgical video analysis, creating a deep learning model to automatically detect and localize surgical instruments and automatically extracting key metrics to assess surgical skill.

Help Received

Serena Yeung, a PhD student at the Stanford Artificial Intelligence Laboratory, answered my questions on how to create the deep learning model, and Dr. Jeffrey Jopling, a surgeon at Stanford Healthcare, provided insight into the metrics that would be most meaningful to analyze.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Sohini Kar

Project Number
S1514

Project Title
Factorization of Recurrence Relations

Abstract

Objectives/Goals
The combinatorial solution to the recurrence relation a_{n} =a_{n-1} +a_{n-3} +a_{n-4} leads to the trend, a_{2n} =f_{n}^2 , where f_{n} is the n-th term in the Fibonacci sequence. This project explores the generalization of the solution to this recurrence relation to yield a new family of sequences where every k-th term is the k-th power of u_{n}: a_{kn} =u_{n}^k. u_{n} is the solution to Horadam#s second order recurrence relation of the kind u_{n}=p#u_{n-1} +q#u_{n-2}, where p, q are integers.

Methods/Materials
I used several techniques to investigate the factorization of these recurrence relations. First, I used diagrams to illustrate the factorization. I explored how to find the recurrence relation with even terms leading to square of generalized second-order recurrence relation through bijections and Binet-like formulas. I took the trend, described above, of squares of Fibonacci numbers for every even term and generalized it for cubes of any second-order sequence. Finally, I derived a generalized version of this sequence, with every k-th term yielding the k-th power of the generalized second-order sequence. Techniques that I used drew from number theory. I used Hadamard products, Cauchy#s residue theorem, diagrams, Binet#s formula, partial fractions, and work by Hoggart and Legendre. An understanding of recurrence relation and generating functions was paramount, as well.

Results
The solution to the recurrence relation was found to be a_{2n} =f_{n}^2 and a_{2n+1} =f_{n}f_{n+1}. The bijection for a_{2n} was denoted by the number of ways we can tile two rectangles of length 1xn with 1x1 square and 1x2 rectangle. This bijection was generalized to k rectangles for a_{kn} and a solution was found for its generating function through bijection as well as Binet-like formula.

Conclusions/Discussion
These results represent the product of a year of investigation, however, additional work is being done to explore related problems in this field, such as examining similar families for Catalan and Motzkin numbers.

Summary Statement
This project derives with the recurrence relation, generating function and bijection for a new family of sequences, where the k-th term a_{kn} =u_{n}^k, where u_{n} is the n-th term of Horadam#s generalized second order sequence.

Help Received
All work on this project was done by me at my home. This project was derived from a problem provided by Dr. Simon Rubinstein-Salzedo and periodically offered input when requested.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

John Kim

Project Number

S1515

Project Title

EmNet: Emotion Recognition from Human Voice Using Machine Learning for Affective Computing

Abstract

Objectives/Goals

The goal of this research is to develop an algorithm that can analyze and accurately recognize the speaker's emotion from human speech using machine learning. This is not only a challenging research problem but also has great potential for real-world applications. Adding human emotions as context information can help transform the emerging human-machine interactions to the next level, especially when video capture of human facial gestures is not a preferable option for the user due to privacy.

Methods/Materials

Conventional approaches to recognize human emotion use pattern recognition on the static features, obtained by taking the mean and variance of time-varying characteristics of feature vectors. However, the procedure to obtain the static features eliminates important information contained in the temporal information of feature vectors. "EmNet," the proposed method in this project, allows the neural network to learn this unknown temporal trajectory. The system consists of feature extraction, followed by Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks.

The data used for this project comes from the Berlin Emotional Speech Database (EMO-DB). The database consists of 535 utterances from 10 talkers, with each utterance representing one of seven different emotional states: anger, boredom, disgust, fear, happiness, and sadness.

Results

The recognition rate achieved by the proposed method reaches above 86%, which is much higher than the 77.3% obtained from the conventional method using Support Vector Machine (SVM). This is a significant error rate reduction by about 40% over the conventional approach.

Conclusions/Discussion

A new method is proposed to recognize emotion from human speech. The method consists of 1) acoustic speech analysis to maximize the efficiency of well-known feature extraction and 2) machine learning to learn an unknown mechanism of temporal information processing.
Validation on other databases and languages remains as further work.

Summary Statement

EmNet is the proposed method for emotion recognition from human speech, comprised of acoustic feature extraction and machine learning, and demonstrates an error rate reduction of about 40% compared to the static approach.

Help Received

Received help to get appropriate packages for feature extraction.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Rachana Madhukara

Project Number

S1516

Project Title

Dynamics on the Rado Graph

Abstract

Objectives/Goals

In this project, we prove certain properties about the Rado Graph. In order to do that, we first consider a simpler case of the graph, a modified version, and prove that it's connected.

Methods/Materials

No materials other than pencils, paper and a computer were needed to conduct this project. That is because this project is a mathematics project.

Results

First, I proved that the Rado graph is connected through the usage of prime gaps and a theorem from Baker, Harman, Pintz. I showed that the graph can keep on growing to infinity, but is only increasing (took the derivative). Then, using results and techniques from other papers, I proved using induction that the Frog Model is recurrent on the Rado graph. Recurrence means that the process will eventually return back to the root as we go to infinity.

Conclusions/Discussion

In this project, we proved that the modified version of the Rado graph is locally connected through the use of many techniques. We used the main theorem from a paper by Benjamini and Peres to conclude that there is a prime in every gap that we jump to. Additionally, we used the idea of polynomial versus exponential growth from a Telcs and Wormald paper. We used this and a theorem from Benjamini and Peres to conclude the frog model is recurrent on the modified graph because it is polynomially growing.

Summary Statement

It is shown in this project that the Frog Model is recurrent on the Rado Graph, a random graph, through the usage of many techniques such as the theory of prime gaps.

Help Received

I did my work in collaboration with Dr. Simon Rubinstein-Salzedo who lives in the Bay Area. He provided tremendous support for me during this project.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Marat Mustafaev; Daniel Rostamloo

Project Number

S1517

Project Title

Sustainability of Katsuwonus pelamis in California Pelagic Fisheries: Analysis of Natural and Fishing Mortality Data

Abstract

Objectives/Goals

The Skipjack Tuna (Katsuwonus Pelamis) is a pelagic species which is currently farmed in California by numerous fisheries. The objective of the project was to determine whether the continued harvest of the species was sustainable.

Methods/Materials

Used data from the National Marine Fisheries Service to represent annual landing and pricing data of the Skipjack Tuna. Microsoft Excel graphing software was used in graphing the data. Mortality data -- both fishing and natural -- were obtained from research papers published by Mark Maunder. A mathematical model which requires the recruitment into adulthood to be greater than or equal to the deaths of the adult population ultimately determined the sustainability.

Results

The mathematical model reflects that the Skipjack Tuna is sustainable. This result is congruent with our earlier hypothesis conjectured using information on fishery restriction in the twentieth century.

Conclusions/Discussion

The model is a highly transferrable utility in measuring the sustainability of other farmed species, fish or otherwise. This is due to the formula's use of simple, easily obtained data sets regarding a harvested species.

Summary Statement

A mathematical model was developed in order to understand and determine the sustainability of farmed species and the Skipjack Tuna in particular.

Help Received

Mark Maunder and the National Marine Fisheries Service for their data on Katsuwonus Pelamis mortality and annual landings, respectively.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Anish R. Neervannan

Project Number

S1518

Project Title

A State-of-the-Art Approach to Detect Diabetic Retinopathy Using Convolutional Neural Networks

Abstract

Objectives/Goals

The purpose of this project was to determine if a state-of-the-art convolutional neural network image recognition algorithm could be developed to detect diabetic retinopathy in Fluorescein Angiography (fudus) photography to match the accuracy of a medical professional's mental model.

Methods/Materials

Materials included a Macbook with high computing power and internet access; a dataset of 35,000 retinal scans from Kaggle's diabetic retinopathy homepage; Inception v3 and ImageNet models; and machine learning libraries such as Tensorflow, Keras, scikit-learn, and matplotlib. A Java program was written to sort the scans based on their pre-classified labels. Python programs were written using the scikit-learn and matplotlib libraries to crop, reshape, denoise, visualize, and transpose the scans. The scans were split into a training set, with 29000 scans, a validation set, with 5000 scans, and a test set, with 200 scans. The Keras and Tensorflow libraries were used to build the Convolutional Neural Network architecture on top of the Inception base model and train the algorithm. The retinal scans in the test set were passed into the trained model and the algorithm's predictions were recorded for each image. The same set of scans was sent to three medical professionals.

Results

After the training process, a controlled test set of 200 scans was used to determine the accuracy, precision scores, and recall scores of the algorithm and the three medical professionals. The image recognition algorithm's accuracy was 79.50%, while the medical professionals' accuracies were 88.00%, 89.00%, and 88.00%.

Conclusions/Discussion

Although the computer algorithm's accuracy came close to the medical professionals' accuracies, it did not match them. The algorithm predicted scans without the disease more consistently than those with the disease. This algorithm breaks new grounds in this field by outperforming a prior model that was published in a Stanford PhD paper (trained on the same image dataset) by an 8% margin.

Summary Statement

I trained a convolutional neural network to detect diabetic retinopathy to be almost as accurate as a medical professional and better than a prior model published in a Stanford PhD paper.

Help Received

Dr. Kapil Kapoor (Vitreoretinal Surgeon), Dr. David Chia (Ophthalmologist), and Dr. Sanjay Kedhar (Gavin Herbert Eye Institute) classified the images in the test set. My father, Raj Neervannan, helped me understand the concepts required to complete the project.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Raj Palleti; Suhas Prasad

Project Number

S1519

Project Title

What Factors Determine an Eulerian Polygon? A Computer-Inspired Analysis

Abstract

Objectives/Goals

Generalize the Euler Line to non-triangular polygons.

Invent a method to find the special points (while confirming collinearity and 2:1 ratio).

Determine which polygons contain the Euler Line.

Methods/Materials

As an initial platform for gaining insight, we began drawing simple sketches to discover possible ways of defining the special points. After realizing the inaccuracy and inefficiency of this approach, we decided to implement the method of computer inspiration. We wrote a program in Java, and simulated random convex polygons using JavaFX and StdDraw (Standard GUIs in Java). In addition, we used the Java Topology Suite (an open source java software library that assists with computational geometry) to calculate the location and ratio of the special points. This allowed us to maintain precision and efficiently gain insight for analyzing an Eulerian Polygon.

Results

Our new method to find the special points of an Eulerian Polygon resulted in three unique points derived from recursively decomposing the polygon into triangles. Using our induction-based recursive formula, we were able to illustrate which polygons were Eulerian.

Conclusions/Discussion

Ultimately, we were able to find the defining criteria of an Eulerian Polygon. Through recursive decomposition into triangles, we could pinpoint the three unique special points of certain polygons. This inductive proof was inspired and not assisted by the use of a computer program, since the program provided insight into developing a mathematical proof and did not offer parts of the proof itself. Upon developing the Euler sequence, we noticed deep isomorphisms to renowned mathematical wonders such as the Collatz conjecture and Fibonacci Sequence, since both rely on a recursive approach to determine the next element. Though already explored to a certain extent, we also wish to extend our novel ideas into the third dimension and explore its practical relation to origami. We are currently working towards publishing a joint arXiv paper explaining our generalizations and contributions.

Summary Statement

Using computer simulations to provide insight, we generalized the Euler Line to all polygons, and discovered a recursive method based on inductive reasoning to determine which polygons are Eulerian.

Help Received

No outside help. From developing our research idea to making our conclusion, all work was collaboratively accomplished between us. During our research, however, we reviewed several online mathematical articles to comprehend existing work.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Austin A. Patel

Project Number

S1520

Project Title

Neural Networks for Handwritten Letter Recognition

Abstract

Objectives/Goals

Teaching a computer to learn to classify handwritten letters requires a complex computing model capable of learning the structure of letters rather than simply following a set of explicitly coded parameters. This challenge requires a machine learning algorithm, such as a neural network that simulates the human brain, to learn over time from training data. The goal was to create a neural network and test various learning algorithms to see which was best for predicting handwritten letters.

Methods/Materials

Materials: Computer with Java programming language installed. A Java program was written from scratch including a user interface for drawing letters, a neural network and three learning algorithms. Fourteen participants drew a total of 3,640 letters used for training and testing. The letters were used to train a neural network and were later classified by the program. Different learning algorithms were used to classify the letters and their performance was analyzed.

Results

By initially training with a genetic then a backpropagation algorithm, the accuracy of predictions reached 80.12%. Most of the algorithms ended up with similar overall accuracies, except for the genetic algorithm which peaked in performance and then got worse. Using multiple algorithms in conjunction resulted in better performance.

Conclusions/Discussion

The results support the idea that algorithms that do not rely on training data are better at predicting letters that do not necessarily align to a specific shape. As detailed in the discussion section, the most significant findings from this experiment were the specific use cases for each algorithm and why each algorithm performs and produces the output it does.

Summary Statement

I programmed a neural network and its learning algorithms that I used to classify what letters people drew.

Help Received

None, I programmed everything and executed the project myself.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Laura C. Pierson

Project Number

S1521

Project Title

Signatures of Stable Multiplicity Spaces in Restrictions of Representations of Symmetric Groups

Abstract

Objectives/Goals

My goal was to find formulas for norms of orthogonal basis vectors in the multiplicity space of copies of an irreducible representation of S_{n-k} in the restriction of an irreducible representation of S_n, and for signatures counting how often the norms are negative.

Methods/Materials

I constructed a stable sequence of multiplicity spaces with fixed k and varying n. I used a basis of standard Young tableaux and applied a series of adjacent transpositions to their entries to find norm formulas in terms of n, then interpolated to arbitrary real values of n and studied the resulting signatures.

Results

I found explicit norm formulas, which give an immediate algorithm for computing signatures. I found that the norms are rational functions of n with positive integer roots coming in pairs that differ by 1, implying that the norm is always positive for sufficiently large or small n. I also found that the norm is the same for all basis vectors in the case where the partitions corresponding to the representation of S_n and S_{n-k} have the same first element. I also found explicit signature formulas in certain cases.

Conclusions/Discussion

The study of stability in representation theory by generalizing to arbitrary real or complex rank is fairly new and leads to rich but largely unexplored structures. My results give interesting information about certain combinatorial invariants in Deligne categories (generalizations of the representations of S_n to an arbitrary complex number), and they help give a better sense of how these categories behave, methods for studying them, and what related formulas might look like.

Summary Statement

I studied generalizations of the representations of the symmetric groups S_n to an arbitrary real number n, which gives richer structures than positive integer cases and provides information about stable properties of the representations.

Help Received

I was matched with the project by the PRIMES-USA program. The problem was suggested by Professor Pavel Etingof (MIT). Siddharth Venkatesh (MIT) helped me understand the necessary background and gave me advice throught the project. I found all the formulas and results myself.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Meghana Reddy; Rakesh Reddy

Project Number

S1522

Project Title

A Corollary to the Pythagorean Theorem: Using Simple Algorithms to Predict Pythagorean Triples

Abstract

Objectives/Goals

The goal of this study was to find a better way to determine all Pythagorean Triples, and test them using computer code written using these equations. Current methods to determine Triples are quite cumbersome and complex, and require 2 arbitrary numbers to determine 3 associated numbers called Pythagorean triples.

Methods/Materials

Started with known, small integer numbers to develop simple equations to show the Pythagorean relationship, and the relationship of numbers within EACH set of Triples. Then, these equations were re-arranged to display a possible trend for solving larger numbers. This lead to developing a NOVEL and unique equation that generated ALL possible Pythagorean Triples. Many have called this a breakthrough of sorts in number theory. All previous methods required 2 arbitrary numbers to generate one or more sets of triples. Our method requires ONLY one number to generate the triples, not one set, but all sets of Pythagorean triples for that number.

Results

The NOVEL equation thus developed yielded four surprising results. (1) It successfully generated solutions for any and all large numbers (2) It produced not one, but ALL possible Pythagorean Triples for a given number (3) It REVEALED a way to determine the maximum possible triples that could be generated starting from a single given value (4) It also generated non-integer solutions that met the Pythagorean equation, which was easily eliminated using the computer code.

Conclusions/Discussion

The solution worked for any and all numbers. The computer code was used to eliminate non-integer solutions. This has wide applications in cryptography and cybersecurity.

Summary Statement

We discovered a novel and unique approach in solving a mathematical problem in number theory that generates Pythagorean triples

Help Received

None. We discovered, formulated and tested our own mathematical equations.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Shalin V. Shah

Project Number

S1523

Project Title

Lumos: Automated Smartphone-based Ophthalmic Screening for Glaucoma Using Computer Vision and Deep Learning Algorithms t

Abstract

Objectives/Goals

The objective is to create 2 things: 1) a low-cost lens that attaches to the back of a smartphone and allows laypeople to take a picture of the back of the eye and 2) an algorithm that analyzes that image to automatically identify risk of glaucoma.

Methods/Materials

Used my Macbook pro laptop plus the programming languages Python and Swift to build the algorithm and app. The machine learning library used was TensorFlow and the algorithm was tested on the MESSIDOR, DRIONS, and HRF databases that all contain high quality retinal images.

The 3d printed casing for the lens was created using a 3d printer as well as OnShape to create the CAD file. The different parts of the lens included: mirrors, beam splitters, polarized filters, and paper.

I created and tested both the lens and the algorithm myself.

Results

The lens ended up being only about 30 dollars while ophthalmoscopes and slit lamps can cost from a few hundred to a few thousand. Additionally, the quality of the lens compares to more expensive lenses, despite being low-cost. Lastly, there is no need for dilation of the pupil when using the lens, because the lens can capture high quality images regardless.

The algorithm, when tested on images captured using the Lumos lens after the first clinical study, was able to identify glaucoma risk with 100% accuracy. The Lumos algorithm, when tested on the high quality image databases HRF and DRIONS, had an overall accuracy of 95.7%.

Conclusions/Discussion

Glaucoma is the leading irreversible cause of blindness worldwide. What's even more alarming is that in its early stages, it is asymptomatic. So more than half of the patients who are affected don't know that they have this illness until much irreparable damage has been done to their vision.

In order to better identify patients who would benefit from early treatment, it is necessary for patients to have early detection of this disease. But early detection is difficult. As it stands, not all patients have access to an ophthalmologist for glaucoma screening due to barriers such as finances, distance and time.

Summary Statement

I created a low-cost lens (smartphone attachment) that lets laypeople take a picture of the back of the eye (retina), and an algorithm that analyzes the retinal image to identify risk of glaucoma.

Help Received

I consulted my mentor, a clinician at UCI Medical Center specializing in glaucoma to further understand the disease. But the lens and algorithm were both created entirely by me.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Arjun S. Subramaniam

Project Number

S1524

Project Title

CadML: A New Computational Approach to Optimizing Antibody Affinity for Design of Antibody Therapeutics

Abstract

Objectives/Goals

Antibody therapeutics is a growing field of drugs, involving antibodies that specifically target a pathogen inside the body. Designing such high-affinity antibodies mimics our immune system's response during infection: mutating and selecting antibody variants that bind better to the antigen (target). Yet instead of doing this in vitro, which is inherently random and time-consuming, we could use computational methods to identify mutations on antibodies that increase their affinity to their target. However, predicting the effects of mutations on how well antibodies bind is still an unsolved problem. I propose CadML, a novel, holistic machine learning- based approach.

Methods/Materials

Given an antibody-antigen mutation, CadML incorporates information about the mutation from all levels of protein structure, from the amino acid sequence to structural elements. In addition, I used convolutional neural networks (CNN), a deep learning model, to extract highly informative features from the surrounding environment of the mutation site. All of this information was combined into a final machine learning model, which predicts the change in binding energy caused by the mutation (measuring binding affinity). CadML was written in Python2.7, with protein modeling tools and Python libraries.

Results

I evaluated CadML on the experimentally-verified AB-Bind dataset, consisting of 528 mutations and associated binding affinity changes (favorable and harmful). CadML achieved a Pearson's Correlation of 0.64, identifying favorable, affinity-improving mutations with high accuracy. My method significantly outperformed state-of-the-art prediction methods, including machine learning and simulation-based methods ($P < 0.01$). CadML achieved a correlation of 0.70 on further evaluation of another dataset.

Conclusions/Discussion

CadML combined deep learning with information from all levels of protein structure to find affinity-improving antibody mutations with high accuracy. I extended my analysis by mapping Rituximab, a current antibody therapeutic for leukemia, and finding mutations that could significantly improve its affinity to cancer cells. In addition, I scanned the surface of HIV to provide insight into therapeutic strategies for AIDS. In the future, I would look to integrate CadML into the drug development pipeline, helping bring high-affinity antibody therapies from the benchside to the bedside.

Summary Statement

I built a machine learning model to design antibody therapies that can treat diseases like cancer and HIV by binding effectively to targets inside the body.

Help Received

I worked with Dr. Thomas MacCarthy in the Department of Mathematics and Statistics at Stony Brook University over the summer (Simons Summer Research Program). I proposed the idea and built the algorithm largely independently, with some guidance from Dr. MacCarthy.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Katherine Tian; Katherine F. Zhang

Project Number

S1525

Project Title

A Novel Approach to Seizure Prediction Using Deep Learning

Abstract

Objectives/Goals

Epilepsy is a neurological disorder affecting more than 65 million people world wide. A seizure prediction system can greatly help epileptic patients to avoid potentially dangerous activities or to timely administer medications.

In this project, we propose the use of deep learning methods such as Recurrent Convolutional Neural Networks (R-CNN) to differentiate between the preictal periods, which signal the imminent onset of a seizure, and the interictal periods of epileptic patients based on their intracranial electroencephalograms (iEEG) data.

Methods/Materials

Using the iEEG-recorded brain data on five canines and two human patients from the American Epilepsy Society Seizure Prediction Challenge 2014, we first created an image-based representation of the multi-channel iEEG signals. Then a recurrent convolutional neural network was trained to capture the spectral, temporal, and spatial patterns in iEEG recordings to predict whether or not the image belongs to the preictal class. In order to find the optimal network structure and parameters, we used the training data to experiment with different network configurations and parameters on Linux/MacOS machines with GPUs, using open source deep learning software packages such as Theano and Lasagne. Cross validation on prediction accuracy was used to evaluate our model's effectiveness.

Results

Our experiments showed that we can predict seizure with 88% accuracy on data from the dogs, and 72.5% accuracy on data from the human patients. This demonstrates that recurrent convolutional neural networks can be effective in seizure prediction.

Conclusions/Discussion

Recurrent convolutional neural networks can be utilized to predict seizure using iEEG data, thus helping epileptic patients live a better and safer life.

Seizure prediction is still an active area of research. Compared to most previous machine learning approaches, ours avoids manually crafting learning features. To our knowledge, there has been no published work that uses a recurrent convolutional neural network in seizure prediction (vs. detection).

Summary Statement

Using a deep recurrent convolutional neural network, we successfully predicted seizure with 88% accuracy on dogs and 72.5% on human patients.

Help Received

We thank Dr. Casso from The Harker School for being our faculty sponsor, and The Harker School for general support. We got our idea from two papers by the following two groups of researchers:1. Thodorof, Pineau, and Lim; 2. Bashivan, Rish, Yeasin, and Codella.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Stanley J. Wang

Project Number

S1526

Project Title

Lost and Found: The Math behind Search and Detection

Abstract

Objectives/Goals

This project uses mathematical concepts, specifically theoretical probability, to create an optimal search model for a target. Computer programs in C++ are also used to perform simulations and calculations.

Methods/Materials

I wrote my own code in C++ to perform sample calculations for simulations.

Results

We have created an optimal search and detection model using probability theory and other mathematical concepts.

Conclusions/Discussion

An optimal search strategy has been formulated, and can be used to find a lost target in almost any situation. This includes, but is not limited to, the work of rescue teams, the police, and the Navy.

Summary Statement

I have developed an optimal search strategy, using both a mathematical method and with implementation in a C++.

Help Received

Professor Atkinson from the Operations Research Department at the Naval Postgraduate School reviewed my results and provided his feedback.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Patrick I. Wildenhain

Project Number

S1527

Project Title

Mitigating Flood Risk by Predicting River Gauge Heights Using a Neural Network

Abstract

Objectives/Goals

The objective of this experiment was to apply artificial neural networks (ANN) to the prediction of river gauge height two days in the future at a specific station in order to mitigate flood risks and damages. As the current method of flood prediction in the US involves a complex mathematical model, a forecast method involving neural networks could offer an extremely important cost effective alternative to not only the US, but also to developing countries around the world.

Methods/Materials

A java program was developed to prepare input data for three different ANNs for assessing the usefulness of specific types of data. All three were given river and precipitation data, however the first ANN was given additional precipitation forecast data. The second ANN was given additional soil moisture data, and the third ANN was given all of the data types. The best network was compared with the National Weather Service forecast (NWS) model.

Results

Soil moisture was found to be essential to predictions, and including predicted precipitation was found to make a statistically significant improvement in the predictions. The best performing network was compared with the NWS model for the year of 2016. The NWS model reached an accuracy of 97.8%, and the ANN model reached an accuracy of 96.6%.

Conclusions/Discussion

This project supplemented traditional forms of data for the training of neural networks for flood forecasting with soil moisture data and precipitation forecast data, and found that both proved to be valuable. In comparing with a current model, the NWS model, the ANN model reached a comparable level of accuracy. While the NWS model requires vast amounts of processing power and data, the ANN model can predict with comparably less processing power and data. Thus, in countries or areas where a supercomputer cannot be afforded, or the vast amounts of data for a mathematical model are not available, the neural network serves as a very important cost effective alternative.

Summary Statement

This project succeeded in utilizing artificial neural networks for flood forecasting, overall managing to predict future water levels with a 96.6% accuracy rate for the year of 2016.

Help Received

Parent helped with code review of the data parsing program.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Aaron S. Willis

Project Number
S1528

Project Title
Scrambling Stats: The Ability of Various Card Shuffling Methods to Produce Logically Arbitrary Card Arrangements

Abstract

Objectives/Goals
If compared against the 3 other most commonly used shuffling methods, the Block shuffle will produce the most random arrangement of cards.

Methods/Materials
Materials used included a computerized pseudorandom card-shuffling program and a half deck of standard playing cards. Using half of a deck expedited the shuffling process and still maintained applicability of the results. Shuffles studied included the "riffle-shuffle", the "block" shuffle, and "smooshing". To set up each shuffle, the cards were rearranged to the same position each time. Card position (measured relative to the top of the deck), trial number, and card was recorded for each trial. The "block" and "riffle" shuffling methods were tested by shuffling 3 times per trial, and the "smooshing" was performed for 30 seconds, the average time it took to perform the other methods.

Results
The shuffling methods did differ quite dramatically from one another, particularly the "riffle" shuffle, which was the least random. The "smooshing" shuffle however, was the most random.

Conclusions/Discussion
The "smooshing" method was shown to be more random than any other shuffles. The "riffle" shuffle was the most volatile among all shuffling method by the metric of average card position. Further, the "riffle" shuffle had many more consecutive card pairs and repeat card positions than the other methods. This implies that the riffle is the most predictable shuffle and that "smooshing" is the least so. Such results suggest that casinos should switch from the traditional "riffle" shuffle to something more able to produce fair results for the players.

Summary Statement
Three methods of card shuffling were compared for their ability to produce random results, and the "smooshing" shuffle was shown to be the best.

Help Received
None. The experiments were designed and performed solely by myself.

CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)
Michelle C. Xu

Project Number
S1529

Project Title
A Novel Approach for Solving Target Mutation-Induced Drug Resistance for HIV-1 Fusion Inhibitors with the Hopfield Neura

Abstract

Objectives/Goals
The formation of a hairpin core structure by the HIV-1 virus transmembrane glycoprotein gp41 is the critical event that triggers viral fusion to the host cell. Fusion inhibitors such as T20 can prevent the formation of a hairpin core by binding with the gp41 N-terminal Heptad Repeat (NHR). However, mutations on the NHR can reduce the binding potency of many inhibitors and can cause HIV-1 drug resistance. This project proposes a novel approach to investigate the binding potencies between the gp41 NHR and various inhibitors, as well as potency loss due to NHR mutations.

Methods/Materials
HIV-1 inhibitors prevent the interaction between the gp41 NHR and CHR that forms the gp41 hairpin core, an essential structure that must be formed in order for a virus to enter a cell. However, mutations on the NHR can affect the potency of inhibitor binding due to changes in molecular interactions. In this project, a hidden Hopfield neural network was identified which can accurately describe the interactions between certain amino acids on the NHR and CHR. The energy model provided by the Hopfield neural network was then applied to study the stability of the hairpin core.
The Hopfield energy model was trained using hydrophobicity scale values from the experimental work from Wimley and White. The energy states (and thus the stability of the structure) of a non-mutated complex and mutated complex were then compared.

Results
Using the Hopfield energy model, the loss of stability of the NHR-inhibitor complex for different mutations and inhibitors was calculated. Results showed that depending on the location of the NHR mutation, some fusion inhibitors will lose their potency while others will still be effective against HIV-1 entry. Results were validated from experimental data.

Conclusions/Discussion
The Hopfield neural network was able to assess the energy stability of the NHR-inhibitor complex. This approach provides a fast way to accurately identify which inhibitor will still be effective against HIV-1 entry based on the mutation. This approach can test for the effectiveness of a new drug before the drug is actually created.

Summary Statement
A hidden Hopfield neural network was identified based on the interactions within the hairpin core, and was applied to study the loss of stability once a mutation occurs.

Help Received
Professor Stephen H. White from the Dept. of Physiology and Biophysics at the University of California, Irvine, and my math teacher Mr. Charles Y. Beilin mentored me on research and answered any questions I had.