DATA 255
Group Project Proposal

Proposal Due date: October 10th, 2023

Project Proposal Approval: email to simon.shim@sjsu.eud and ‘cc to supreetha.naik@sjsu.edu and neha.shaikh@sjsu.edu
competition title if it is from Kaggle

- One paragraph abstract
- Dataset links and description
- Technologies proposed to use like Python, Tensorflow, keras etc.
- Reference links

Your solution must include models using deep learning.

If you are going to choose a Kaggle competition (https://www.kaggle.com/competitions), choose **an active competition** (with due date after 2 months). Do not choose Completed ones unless you can make significant changes in the problem requirements.

What is Kaggle? Kaggle is an online community platform for data scientists and machine learning enthusiasts. Do not ignore/underestimate Kaggle profile.
How to compete at a Kaggle contest?
Try to handle multiple versions of models (multiple models)
Try to handle multiple versions of datasets(engineer the existing datasets)
Learn to tune hyperparameters efficiently
Try to validate effectively

Steps to follow:

1. **start with a relatively simple model, don’t jump into most complex model first**
2. **Vary one thing at a time and record history and reasoning.**
3. **Try to test multiple ideas concurrently**
4. **Fail fast and restart**

Definitely, a team has advantages

Other project ideas:

- **SIEM**- security information and events management, where UBA is a critical part. UBA- user and entity behavioral analytics. Unsupervised ML to find threats for insider threats. This is important to companies because insider threats are very hard to predict. What behavior are you doing, do you have a virus/hacker? You have some sensitive data with these data to detect hacks but that also means they cant get hacked with this sensitive info like social security numbers. Steps: analyze the data- network activity/login attempts, baseline the data- user/dept./region/company behavior, advanced threat detection- malware, blacklisted IP address. UBA works by using the data sources like logs (network, server, identity) to ML model to produce anomaly classifications like suspicious data, flight risk user. 3-5 threats a day- cant investigate everything.

  Measuring abstract reasoning in neural networks

  Whether neural networks can learn abstract reasoning or whether they merely rely on superficial statistics is a topic of recent debate. Here, we propose a dataset and challenge designed to probe abstract reasoning, inspired by a well-known human IQ test. To succeed at this challenge, models must cope with various generalisation 'regimes' in which the training and test data differ in clearlydefined ways. We show that popular models such as ResNets perform poorly, even when the training and test sets differ only minimally, and we present a novel architecture, with a structure designed to encourage reasoning, that does significantly better. When we vary the way in which the test questions and training data differ, we find that our model is notably proficient at certain forms of generalisation, but notably weak at others. We further show that the model's ability to generalise improves markedly if it is trained to predict symbolic explanations for its answers. Altogether, we introduce and explore ways to both
  measure and induce stronger abstract reasoning in neural networks. Our freely-available dataset should motivate further progress in this direction
- **Active Load Management (ALM) Forecasting**: Minute level has too much noise – want peak ALM, want to serve maximum number of people. Don't need to clean minute level, just use daily level peak. Seasonality is an important thing- week days are different than weekends. Time of day not much difference. Yearly seasonality also has trends. Estimate trends and seasonality. Time series/forecasting extrapolate data which goes against statistics. LSTM doesn't really accommodate for trend, seasonality. Forecasting for profit- Bayesian modeling, superimpose every month year and see where the data is going. Look at seasonality, trend, smoothing for daily levels. Can do hierarchical analysis- global level aggregated, region level continents accessing Facebook in different way (culture, time zones), countries—helps get a better forecast. Make sure all levels of forecast are consistent. If you can forecast regions more precisely- you can do better at global level. Have clear understanding of where your demand is coming from can help set up data centers. Understanding how people are using services- emphasis on certain services for certain time of day. Outliers aren't always clear but its an important part. Done in a cycle- what you think the right value should be based on forecast and run the process again. Special events- average the past 10 years etc. step 1) point-estimate, step 2) prediction/confidence intervals from time series, step 3) bootstrapping on Bayesian side. Now have forecasting- next need to test/keep forecasting- error of new data coming in but data might change over time, doesn't mean forecast was bad. Plus outliers happen.
- **Sequential transfer learning for Information extraction**: Three models were used: BERT (first model built on encoder stack), ALBERT and RoBERTa(better optimization and outperforms BERT). They are pretrained on Wikitext and BookCorpus and then fine tune on Intuit documents and run OCR and have entity labels to extract entity data and thus build custom token classification output. The output of OCR is tokenized, and tokens are converted to vector and are given to the BERT model. The encoder layers in BERT convert this input to inner dimension to be given to the classification layer/data classifier. This labels the tokens to different entity/class. Confidence model is built for the model to return the confidence along with the extraction.

This extraction for one document at a time. Thus, multiple models need to be created for different document type and this takes more memory (~400MB), run time, requires more data, optimize different parameters for each model, difficult to deploy.

- **Healthcare AI use case**: a) Breast cancer screening: Niramai is an Indian based company working on breast cancer screening without mammography. They use thermal images and build segmentation models (like RPM) using AI which predicts the malignancy probability for each pixel value. The equipment built is non-invasive and the main advantage of using DL in healthcare is not to replace radiologists but to provide extra eyes to highlight affected areas.
  b) Classification of Melanoma: Google's Inceptionv3 CNN pre-trained model is used to detect skin cancer using images of the skin. The CNN model was on par with the 21 tested experts to classify skin cancer. The model is portable and easy to use for the patients where we just upload the image of our skin and the model predicts the skin cancer with the malignant probability chances. The model classifies 757 individual classes.
  c) Classification of Pneumonia: CheXNet model is a 121-layer CNN model use to classify pneumonia. This model takes the images of the chest and the results were compared with that of the radiologists. The F1 score showed the AI model (0.435) performed better than the radiologists (0.38).Predicting Kidney Failure: The model uses RNN to predict the kidney failure using medical records of the patients. The results showed the model predicted the kidney injury 48 hours prior compared to existing diagnosis and predicted correctly for 9/10 patients.
  d) Drug Discovery: Molecular structures are used for drug discovery using graph neural net with no fixed structure. The input drug data is labeled, and the model can look at the structure of the drug during training and this is used on the unknown drug. This saves lot of money, time and human effort.

5. **Ubiquitous cyber-intrusions** endanger the security of our devices constantly. They may bring irreversible damages to the system and cause leakage of privacy. Thus, Intrusion detection systems (IDS), as one of the important security solutions, are used to detect network attacks. With the extensive applications of traditional machine learning algorithms in the security field, intrusion detection methods based on machine learning techniques have been developed rapidly. However, Intrusion Detection Systems are weak against adversarial attacks, and research is being done to prove the ease of breaking these systems. To improve the detection system, more potential attack approaches should be researched. Our goal in this project is to design a framework called Fast - Intrusion Detection System Generative Adversarial Network (F-IDSGAN) to create adversarial attacks, which can deceive and evade any IDS. Based on the CICIDS2017 dataset, the developed model is able to generate different attacks and excellent results are achieved.

Deep Entity Classification:

Abusive Account Detection for Online Social Networks Online social networks (OSNs) attract attackers that use abusive accounts to conduct malicious activities
for economic, political, and personal gain. In response, OSNs often deploy abusive account classifiers using machine learning (ML) approaches. However, a practical, effective ML-based defense requires carefully engineering features that are robust to adversarial manipulation, obtaining enough ground
truth labeled data for model training, and designing a system that can scale to all active accounts on an OSN (potentially in the billions). To address these challenges, we present Deep Entity Classification (DEC), an ML framework that detects abusive accounts in OSNs that have evaded other, traditional abuse detection systems. We leverage the insight that while accounts in isolation may be difficult to classify, their embeddings in the social graph—the network structure,

properties, and behaviors of themselves and those around them—are fundamentally difficult for attackers to replicate or manipulate

A deep learning system for differential diagnosis of skin diseases

Skin conditions affect 1.9 billion people. Because of a shortage of dermatologists, most cases are seen instead by general practitioners with lower diagnostic accuracy. We present a deep learning system (DLS) to provide a differential diagnosis of skin conditions using 16,114 de-identified cases (photographs and clinical data) from a tele-dermatology practice serving 17 sites. The DLS distinguishes between 26 common skin conditions, representing 80% of cases seen in primary care, while also providing a secondary prediction covering 419 skin conditions. On 963 validation cases, where a rotating panel of three board-certified dermatologists defined the reference standard, the DLS was non-inferior to six other dermatologists and superior to six primary care physicians (PCPs) and six nurse practitioners (NPs) (top-1 accuracy: 0.66 DLS, 0.63 dermatologists, 0.44 PCPs and 0.40 NPs). These results highlight the potential of the DLS to assist general practitioners in diagnosing skin conditions.

Submit a final report of your project with the following:

* Literature review/related work (5 pts)

* Project Architecture (5 pts)

*models tried and results (10 pts)

* rankings, conclusion, future work (15 pts)