King Fahd University of Petroleum & Minerals
College of Computing and Mathematics
Computer Engineering Department

COE 292: Introduction to Artificial Intelligence

Class Project

General information

Your project can address any given problem that utilizes one of the known Artificial Intelligence (AI) algorithms covered in the course to solve some issue. Particularly, the project should consist of an implementation of a machine learning algorithm or neural network to solve a given problem. Your solution should include explanation about the used technique, analysis and possible improvement of some existing AI system or technique. The main goal of the project is demonstrating the student's ability to effectively develop an AI software and enhance its performance.

You must select one of the projects listed in this document and inform your instructor accordingly as per the deadlines specified. If you choose to select another topic, kindly contact your instructor, and agree with him on the topic and deliverables.

Group Formation Rules

You should work in a small group consisting of a maximum of 3 students. Your instructors will assign a group for you, and you will be required to collaborate with all members of the group to fulfil the project requirement.

Evaluation of the Team Members

The group report will have to identify what part of the project was done by each team member and the amount of contribution made by each member. Each team member will be asked to evaluate his team members and assign a grade to them. This will have direct impact on the grade they will obtain.

Deliverables

1. Proposal Report (report 1)
2. Progress Report (report 2)
3. Final Report
4. 10-minutes video

Project Process

To demonstrate mastery of your skills, each team will be required to deliver two mini reports and a final report. In the first mini report, the team will need to identify the problem they wish to solve and what method of AI they think is best suited for the project and why. In the second report the team must show that they have obtained, cleaned, parsed and scaled the data to make it ready for the AI algorithm. This step should also filter outliers and erroneous data if any. In the final report, the team will show the output of their classification techniques and how it performs.

Each team will require to deliver with the final report a 10-minute video where the video details the work done in the project and the share of each team member in it. There will be assigned dates where team members will be required to play the video of their work in class and then justify to the class and the instructor their findings. Note that each team member should be given an equal participating share in the video and when asked question about their part, they should be capable of answering the question. Be creative in making the videos as you will play them in front of the class and the best video may be used by the department to demonstrate the output of the course.

Evaluation Criteria

You are to submit three reports and a video in the duration of the project. The details are as follows:

1. Report 1: a one-page proposal that should identify the following: (10 points)
    a. The chosen Topic
    b. The chosen AI algorithm and why did you choose this algorithm to solve this problem.
    c. Project plan where overall project tasks are clearly identified and assigned to group members.
2. Report 2: expand your proposal and identify the project progress. The following should be discussed: (20 points)
    a. Obtaining the data and parsing the data set (5 points)
    b. Cleaning and conditioning the data such as scaling removing outlier …etc. (5 points)
    c. Technical task division between team members. Improve on the project plan and divide the technical tasks among team members. (5 points)
    d. Expected enhancement/achievements in the project. (5 points)
3. Final Report: the report should have the following: (70 points)
    a. A short background on the used AI algorithm (10 points)
    b. Result of using AI and discussion (10 points)
    c. What could have been done better to improve the result. (10 points)
    d. A 10-minute video presentation discussing the problem, the AI algorithm used, and the results found. Stress on how you enhanced a pre-trained AI algorithm if you used one or how your output is unique. (20 points)
    e. Discussion and question answering (10 points)
    f. Peer evaluation (10 points)

Due Dates

Report 1 is due on 13th November 2021
Report 2 is due on 27th November 2021
Final report and video presentation is due on 14th December 2021
Your instructor will set a date in class for your team to present in the final week.

Example of Projects:

Project 1: Resume Parser

Recruiters spend a lot of time skimming through resumes to find the best candidate for a job position. Since there can be hundreds of applications for a single position, this process has been automated in several ways where the most common is keyword matching. Resumes are shortlisted and read by the recruiters based on a set of keywords found in a candidates resume. Otherwise, the resume is discarded, and the candidate is rejected for the job. However, this screening process has many drawbacks. Candidates are aware of the keyword matching algorithm, and many of them insert as many keywords as possible into their resumes to get shortlisted by the company.

You can build a resume parser with the help of artificial intelligence and machine learning techniques that can skim through a candidate's application and identify skilled candidates, filtering out people who fill their resume with unnecessary keywords.

You can use the Resume Dataset available on Kaggle to build this model. This dataset contains only two columns, (1) job title and (2) the candidate's resume information.

The data is present in the form of text and needs to be pre-processed. You can use the NLTK Python library for this purpose. Then, you can build a clustering algorithm that groups closely related words and skills that a candidate should possess in each domain. Words that are similar in context (and not just keywords) should be considered. You can assign a final weightage score to each resume from 0 (least favorable) to 10 (most favorable).

Dataset: Kaggle Resume Dataset

Project 2: Translator App

If you are interested in getting started in the field of Natural Language Processing, you should try building a translator app with the help of a transformer. A transformer model extracts features from sentences and determines the importance of each word in a sentence. A transformer has an encoding and decoding component, both of which are trained end-to-end.

You can build your own AI translator app with a transformer. To do this, you can load a pre-trained transformer model into Python. Then, transform the text you want to translate into tokens and feed it into the pre-trained model.

You can use the GluonNLP library for this purpose. You can also load the train and test dataset for this AI project from this library.

Python Package: GluonNLP

Project 3: Object Detection System

You can demonstrate skills in the field of computer vision with this project. An object detection system can identify classes of objects present within an image.

For example, suppose an image contains a picture of you working on a laptop. In that case, an object detection system should be able to identify and label you (human) and the computer, along with your position in the image.

You can use Kaggle's Open Images Object Detection dataset for this project. There is a pre-trained object detection model that has been made open source called SSD. This model was trained on a dataset of everyday objects called COCO and can identify things like tables, chairs, and books.

You can further train the output layer of this model on the Kaggle Open Images dataset to build your object detection system with high accuracy.

Dataset: Kaggle Open Images Object Detection Dataset

Project 4: Animal Species Prediction

Another interesting computer vision project you can do is to predict an animal's species based on an image. You can do this with the Animals-10 dataset on Kaggle. There are ten different categories of animals in this dataset dog, cat, horse, spider, butterfly, chicken, sheep, cow, squirrel, elephant. This is a multi-class classification problem, and you will need to predict the species of the animal based on its picture in the dataset.

You can use a pre-trained model called VGG-16 for this purpose. You can load this model into Python with the Keras library. VGG-16 is a Convolution Neural Net (CNN) architecture trained on ImageNet, which contains over 14 million images. It consists of pictures of everyday objects, fruits, vehicles, and certain species of animals. After loading the VGG-16 model into Python, you can train on top of it with the labelled images in the Kaggle dataset to classify the ten different types of animals.

Dataset: Animals-10 Kaggle Dataset

Project 5: Pneumonia Detection with Python

Many diseases such as cancer, tumors, and pneumonia are detected using computer-aided diagnosis with the help of AI models.

There are open image datasets available on Kaggle for disease detection. You can try your hand with disease prediction on one of these datasets the Chest X-Ray Images (Pneumonia Detection) dataset on Kaggle. This dataset consists of three types of labelled lung X-Ray images: Normal, Bacterial Pneumonia, and Viral Pneumonia. You can build a model that categorizes a patient's health condition into one of these three categories based on an X-Ray image of their lungs.

To build this model, you can use a Python library called FastAI. FastAI is an open-source library that allows users to quickly create and train deep learning models for various problems, including computer vision and NLP. This library provides a higher level of abstraction than Keras and is very easy to work with if you are a beginner. A problem that takes over 30 lines to solve with Keras can be solved in only five lines of code with FastAI.

You can download the ResNet50 pre-trained model from FastAI and train on top of this model to build the classifier. ResNet50 allows us to train incredibly deep neural networks with over 150 layers, and training on top of it will give you good results.

Dataset: Kaggle Chest X-Ray Images

Project 6: Predict Stock Prices

The stock market is like candy-land for any data scientists who are even remotely interested in finance. First, you have many types of data that you can choose from. You can find prices, fundamentals, global macroeconomic indicators, volatility indices, …etc., and the list goes on and on. Second, the data can be very granular. You can easily get time series data by day (or even minute) for each company, which allows you think creatively about trading strategies. Finally, the financial markets generally have short feedback cycles. Therefore, you can quickly validate your predictions on new data. Some examples of beginner-friendly machine learning projects you could try include…

- **Quantitative value investing…** Predict 6-month price movements based fundamental indicators from companies' quarterly reports.
- **Forecasting…** Build time series models, or even recurrent neural networks, on the delta between implied and actual volatility.
- **Statistical arbitrage…** Find similar stocks based on their price movements and other factors and look for periods when their prices diverge.

Tutorials

- Python: sklearn for Investing – YouTube video series on applying machine learning to investing.

Data Sources

- Quandl – Data market that provides free (and premium) financial and economic data. For example, you can bulk download end-of-day stock prices for over 3000 US companies or economic data from the Federal Reserve.
- Quantopian – Quantitative finance community that offers a free platform for developing trading algorithm. Includes datasets.
- US Fundamentals Archive – 5 years of fundamentals data for 5000+ U.S. companies.

Project 7: Teach a Neural Network to Read Handwriting

Neural networks and deep learning are two success stories in modern artificial intelligence. They've led to major advances in image recognition, automatic text generation, and even in self-driving cars.

To get involved with this exciting field, you should start with a manageable dataset.

The **MNIST Handwritten Digit Classification Challenge** is the classic entry point. Image data is generally harder to work with than "flat" relational data. The MNIST data is beginner-friendly and is small enough to fit on one computer.

Handwriting recognition will challenge you, but it doesn't need high computational power.

To start, we recommend with the first chapter in the tutorial below. It will teach you how to build a neural network from scratch that solves the MNIST challenge with high accuracy.

Tutorial

- Neural Networks and Deep Learning (Online Book) – Chapter 1 walks through how to write a neural network from scratch in Python to classify digits from MNIST. The author also gives a very good explanation of the intuition behind neural networks.

Data Sources

- MNIST – MNIST is a modified subset of two datasets collected by the U.S. National Institute of Standards and Technology. It contains 70,000 labeled images of handwritten digits.

Project 8: Predict Housing Price

In this project, you will have to predict the selling price of a new home in Boston. The dataset of this project contains the prices of houses in different areas of the city. You can get the datasets for this project at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set).

Project 9: Improve Health Care

Another industry that's undergoing rapid changes thanks to machine learning is global health and health care. In most countries, becoming a doctor requires many years of education. It's a demanding field with long hours, high stakes, and an even higher barrier to entry. As a result, there has recently been significant effort to alleviate doctors' workload and improve the overall efficiency of the health care system with the help of machine learning. Uses cases include:

- **Preventative care…** Predicting disease outbreaks on both the individual and the community level.
- **Diagnostic care…** Automatically classifying image data, such as scans, x-rays, etc.
- **Insurance…** Adjusting insurance premiums based on publicly available risk factors.

As hospitals continue to modernize patient records and as we collect more granular health data, there will be an influx of low-hanging fruit opportunities for data scientists to make a difference.

Tutorials

- Machine Learning in Health Care – Excellent presentation by Microsoft Research

Data Sources

- Large Health Data Sets – Collection of large health-related datasets
- data.gov/health – Datasets related to health and health care provided by the U.S. government.
- Health Nutrition and Population Statistics – Global health, nutrition, and population statistics provided by the World Bank.

Project 10: Bring your own project

The student may bring his own project and idea and discuss with the instructor. In this case, the student will require to provide the data he will be using by the time of submitting progress report 2. We encourage new project ideas related to your field, related to problems faced by a community or any project you may think of where AI can be used to solve the problem. In this project, you will need the following

1. Approval of your team to work on the project
2. Approval of the instructor to undertake the project in the proposal report (report 1). Please note that extremely hard problems and extremely easy problems will be rejected by the instructor.
3. You will need to provide the data you will use after cleaning and scaling it by report 2.

Project 11: Build AI to play a game

Select a game that you want and build an AI program that can play the game. Note, that games discussed in class for which the code has been given cannot be considered as project. You can go and build any other game that uses advanced search methods to solve the problem. You will need to specify the game and which AI algorithm you plan do the game. The game should not be a trivial game and your advisor will have to approve the game after the submission of your first report.