Group 1
Members: Teng Li, Yifeng Liu, Kaiyue Zheng

Group Project Proposal

Data source: https://www.kaggle.com/c/statoil-iceberg-classifier-challenge

1. **What problem did you select and why did you select it?**

   We selected a competition problem that was posted on Kaggle.com. The name of the competition is **Statoil/C-CORE Iceberg Classifier Challenge** and the goal is to to build an algorithm that automatically identifies if a remotely sensed target is a ship or iceberg. Drifting icebergs present threats to navigation and activities in areas such as offshore of the East Coast of Canada. Currently, many institutions and companies use aerial reconnaissance and shore-based support to monitor environmental conditions and assess risks from icebergs. However, in remote areas with particularly harsh weather, these methods are not feasible, and the only viable monitoring option is via satellite. Statoil, an international energy company operating worldwide, has worked closely with companies like C-CORE. C-CORE have been using satellite data for over 30 years and have built a computer vision based surveillance system. To keep operations safe and efficient, Statoil is interested in getting a fresh new perspective on how to use machine learning to more accurately detect and discriminate against threatening icebergs as early as possible. The solution to this problem will contribute to helping drive the costs down for maintaining safe working conditions.

   We are interested in this problem as it represents a **real world problem,** where you want to obtain a high accuracy to keep operations safe and meanwhile, lower the cost as much as possible. Additionally, it's a problem connected to many domain knowledge in Physics, Aerospace and Geography, which makes the project difficult but also very interesting. From a data science perspective, this problem is quite intriguing as it includes dealing with **image classification, data augmentation** and other challenges that we are looking forward to.

2. **What database/dataset will you use? Is it large enough to train a deep network?**

   There are two data sets that will be used, both offered by this competition:
   a training set and a test set and both of them are presented in json format.The files consist of a list of images, and for each image.

   There are 1604 observations and 4 variables in the training set. To deal with low number of samples in training set, we will do data augmentation to improve our classification accuracy.

3. **What deep network will you use? Will it be a standard form of the network, or will you have to customize it?**

In this project, we are going to use three different deep networks which are MLP, SVM and CNN. Basically, we will build these models depending on some academic papers and tune parameters to get better performance. We will compare each deep networks with their performance to get our best network model.

4. **What framework will you use to implement the network? Why?**

- Pandas: We will use this for data manipulation.
- Scikit-learn: Provides necessary libraries for regression, clustering, preprocessing.
- Pytorch: We are familiar with pytorch and it is easier for us to customize our own network and also to get a better debugging experience.
- TensorFlow: For trying different networks that are best suited for comparison with the well performing base case models. We will use “TensorBoard” to supervise and evaluate the training process.
- Keras: Another option other than Pythorch to build CNN.

5. **What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?**
    a. *Kaggle.com*
    b. *NN Design Book by Hagan, Demuth, Beale, De Jesus*
    c. *Python Machine Learning by Sebastian Raschka*
    d. *Machine Learning in Python by Michael Bowles*

6. **How will you judge the performance of the network? What metrics will you use?**

For local performance we’ll use cross-validation with a log loss metric per the competitions requirements.

For global performance comparison we will be submitting our results to the competition which will gives us a score, and based on that we will judge the performance of our network.

7. **Provide a rough schedule for completing the project.**

Project Action	Dates
Prepare Data: Data Cleaning & Formatting	04.02 - 04.08
Choosing Model	04.09

Training Model, Tuning, Prediction & Analysis	04.10 - 04.17
Model Validation & Evaluation, Last Minute Tuning	04.18 - 04.20
Conclusion & Final Report	04.21 - 04.24