Precipitation Forecast

Problem Definition Framework

The steps I apply in solving the task is by understanding the problem. The challenge is to train a model that take images of daily precipitation maps as input, and generate precipitation forecast maps for one week (7 days) into the future. According to Tom Mitchell, machine learning is a computer program that learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Therefore, this program

1. Data manipulation

For data manipulation; we consolidated the precipitation map file into an array so that they can be used as an index in Pandas and NumPy. We used Python Imaging Library to load the image from a directory, with the open() function in the Image module passing it the path to the image. If successful, the function returns an Image object. The following were the steps I took to process the images:

a. Processing of the Images as object.
b. Resizing Images

2. Choice of model

The choice of model in this task was a straight Convolutional Neural Network (CNN) with future plan to implement the model with CNN LSTM architecture. In this task, I use the CNN layers for feature extraction on input image data combined with sequential to support sequence prediction. We further define the model wrapping each layer in the CNN model with a separate Convolution2D layer. This is to ensure that the model summary provides a clear idea of how the network hangs together. The CNN model is a feature extraction model with hope that the vector output of the Flatten layer is a compressed and/or more salient representation of the image than the raw pixel values.

3. Model parameters

Model parameters are internal to the model and they are the properties of the training data that are learnt during training by the classifier (i.e. learned by the training algorithm) or other machine learning model. For instance in this task, we build a CNN model, a Conv2D as an input layer with 2 filters and a

3x3 kernel to pass across the input images. Model parameters differ for each experiment and depend on the type of data and task at hand. Please refer to the source code.

4. Training algorithm

The step-by-step procedure for adjusting the connection layers of the CNN. In this task, the desired (correct) output for each input image (vector) of a training set is presented to the network, and we iterate through the training datasets. We also adjusted the CNN weights.

5. Evaluation metrics

See source code for documentation

6. Results

See source code for documentation

Appendix A - Bonus Questions

To improve your chances, answer the following questions in your report. You may include these answers as part of discussions throughout your report, or answer them directly in a separate section.

1. Discuss two machine learning approaches that would also be able to perform the task. Explain how these methods could be applied instead of your chosen method.

The two approaches include:

a. Convolutional neural network - Long Short-Term Memory (CNN LSTM): A convolutional neural network is used to learn features in spatial input like images and the LSTM can be used to support a sequence of images as input or generate a sequence in response to an image. LSTM is able to solve many time series tasks unsolvable by feedforward networks using fixed size time windows.

b. Multilayer perceptron neural networks (MLP): The application of MLPs to sequence prediction requires that the input sequence be divided into smaller overlapping subsequence's that are shown to the network in order to generate a prediction. The time steps of the input sequence become input features to the network. The subsequence's are overlapping to simulate a window being slid along the sequence in order to generate the required output.

2. Explain the difference between supervised and unsupervised learning.

Supervised learning is a type of machine learning that is used on a problem where it main aim is to learn a mapping from inputs to outputs. The process of mapping from inputs to outputs are referred to as "*supervised*" because the learning process operates like a Professor supervising a research student. In supervise learning makes predictions, the predictions are then compared to an expected outcomes. Examples of supervised machine learning problems include: **Classification (**the mapping of input variables to a label) and **Regression (**the mapping if input variables to a quantity). Also, examples of supervised machine learning algorithms include: k-nearest neighbours, support vector machines, multilayer perceptron neural networks.

Unsupervised learning is a type of machine learning that is used on a problem where there are only the inputs, and the main goal is to learn the inherent interesting structure in the data. The process to learn the inherent is refer to as "*unsupervised*". This distinguish the method from the "*supervised*" methods. Unsupervised model require no supervision, instead the models are updated based on repeated exposure to examples from the problem domain. Examples of unsupervised machine learning problems include: Clustering (i.e. the learning of the groups in the data) and association of the learning of relationships in the data. More so, an examples of unsupervised machine learning algorithms include: apriori, self-organizing map neural network and k-means.

3. Explain the difference between classification and regression.

Classification involves assigning an observation a label while regression involves predicting a numerical quantity for an observation. Classification and regression predictive modelling problems are two high level types of problems, although there are many specializations, such as recommender systems, time series forecasting and much more.

4. In supervised classification tasks, we are often faced with datasets in which one of the classes is much more prevalent than any of the other classes. This condition is called *class imbalance.* Explain the consequences of class imbalance in the context of machine learning.

Class Imbalanced refers to a problem with classification problems where classes (i.e. Imbalance datasets) are not represented equally. The consequences of class imbalance in the context of machine learning is that imbalance datasets degrades the performance of machine learning

model techniques as well as the overall accuracy and decision making be biased to the majority class, which lead to misclassifying the minority class samples or furthermore treated them as noise.

5. **Explain how any negative consequences of the class imbalance problem (explained in question 4) can be mitigated.**

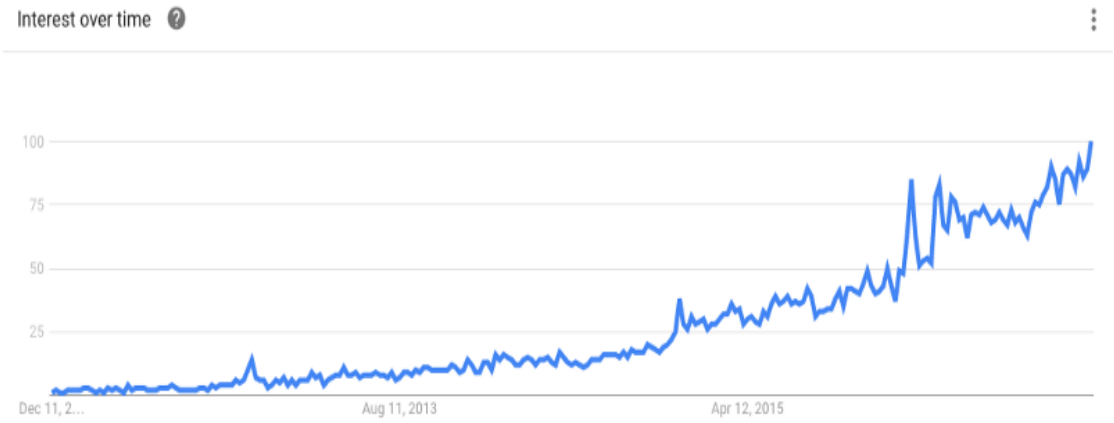
One way of mitigating the problem is to carry out an artificial subsampling or up-sampling in the training set to balance the datasets. Moreover, it is usually better to have a balanced training set, since otherwise the decision boundary is going to give too much space to the bigger class and you are going to misclassify too much the small class. Another way is to use cost based methods, where one weight the importance of every dataset so that the loss function has more contribution from the datasets of the most important class.

6. **Provide a short overview of the key differences between deep learning and any non-deep learning method.**

Deep Learning is a subset of Machine Learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.

7. **What is the most recent development in AI that you are aware of and what is its application, if any?**

The most recent development in artificial intelligence can seem overwhelming, but it really goes down to two very popular concepts Machine Learning and Deep Learning. Although, Deep Learning is gaining much popularity due to its supremacy in terms of accuracy when trained with huge amount of data (see below diagram)



Trend of “Deep Learning” in google

Excerpted from: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

8. Explain how the above development moves our understanding of the field forward.

Understanding of the field allows us to:

a. as organizations expand the use of machine learning for profiling and automated decisions, there is growing concern about the potential for bias
b. draw a distinction between “interpretability by inspection” versus “functional” interpretability.
c. make deep learning practical, and needed a lot of computing horsepower.
d. learn Applied Machine Learning Algorithms

References

1. Applying LSTM to Time Series Predictable through Time-Window Approaches, 2001
2. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic. Modeling, 2014
3. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting: https://arxiv.org/pdf/1506.04214v2.pdf
4. Creating Automatic Gifs with Mask R-CNN: https://www.makeartwithpython.com/blog/automatic-gifs-with-deep-learning/
5. Multivariate Time Series Forecasting with LSTMs in Keras: https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
6. Jian, C., Gao, J. and Ao, Y., 2016. A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, *193*, pp.115-122.
7. Longadge, R. and Dongre, S., 2013. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.
8. Learning from Imbalanced Data: http://www.ele.uri.edu/faculty/he/PDFfiles/ImbalancedLearning.pdf
9. A comprehensive tutorial towards 2D convolution and image filtering: https://github.com/Machinelearninguru/Image-Processing-Computer-Vision/tree/master/basics/Image%20Filtering