基於深度學習技術用於預測海平面高度之變化
(Application of Deep Learning Technology to Predict Changes in Sea Level)

Yi-Lin Hsieh
Department of Data Science, Soochow University, Taipei, Taiwan
lilian790120@gmail.com

Ming-Hsiang Su
Department of Data Science, Soochow University, Taipei, Taiwan
Huntfox.su@gmail.com

摘要

根據世界氣象組織的數據顯示，地球上的氣溫從 1850 年至 2020 年已升高將近 1℃，造成全球氣候異常，嚴重影響南極冰層與格陵蘭冰層融化。若在未來 100 年內，冰層全部完全溶化，海平面就會上升67.2米，大部份沿海城市都會沉沒於海中，四面環海的島嶼國家有可能從此消失。本研究以預測海平面升高之高度為主要探討，由歷年的二氧化碳推算出全球溫度、冰層面積，預測未來海平面將會升高多少。本研究利用全球歷年二氧化碳、全球歷年氣溫、全球歷年北極海冰面積、全球歷年海平面高度資料進行分析，並使用線性回歸與長短期記憶網路交互預測。首先使用線性回歸個別觀察每個數據的指數關係，再使用長短期記憶網路訓練互相有關係的數據，最後使用未來的年份當作預測值，推算每一個未來數值並預測未來 30 年或 50 年的海平面高度。

Abstract

According to data from the World Meteorological Organization, the temperature on Earth has risen by nearly 1°C from 1850 to 2020, causing global climate anomalies and seriously affecting the melting of Antarctic ice and Greenland ice. In the next 100 years, if the ice completely melts, the sea level will rise by 67.2 meters. As a result, most coastal cities will be submerged in the sea, and the island nation surrounded by the sea may disappear. This study focuses on predicting the height of sea level rise, calculates global temperature and ice area from carbon dioxide over the years, and predicts how much sea level will rise in the future. This study uses global historical carbon dioxide, global temperature, global Arctic sea ice extent, global historical sea level data for analysis, and uses linear regression and long short-term memory network interactive prediction. First, this study uses linear regression to individually observe the exponential relationship of each data. Second, this study uses long short-term memory networks to train data that are related to each other. Finally, this study uses future years as forecast values, extrapolates each future value and predicts sea level heights for the next 30 or 50 years.

關鍵字：海平面高度、氣候變遷、線性回歸、長短期記憶網路

Keywords: Sea Level, Climate Change, Linear Regression, Long Short-Term Memory network

1 Introduction

巴黎氣候協定是 195 個國家在 2015 年達成的劃時代協定，設法藉由抑制全球溫室氣體排放量，來避免氣候變遷所產生的嚴重效應。2019 年 11 月 4 日美國正式退出「巴黎氣候協定」，為了使外移工廠願意遷回美國，開放燃煤，使煤礦業、重工業、製造業重新啟動，美國排碳限制的解除，導致聯合國無法掌握未來溫室氣體帶來的氣候變遷。科學家們相信，全球溫度在未來幾十年內將繼續上升，這歸因於人類產生的溫室氣體導致的溫室效應。而二氧化碳是造成溫室效應最主要的氣體，其原因是二氧化碳相對於其他溫室氣體影響溫室效應的時間性更為長久。在大自然中生物與植物也都會自行產生二氧化碳，但根據美國環保局 (EPA) 的統計，人類產生的二氧化碳占全球的溫室氣體 77% (EPA, 2022)。除了人類所產生的二氧化碳，近期的森林大火和火山爆發也造成二氧化碳遽增，植物的

減少加上人類的過度發展，使二氧化碳循環平衡破壞，造成二氧化碳每年加速累積，經紅外線輻射吸收留住能量，導致全球表面溫度升高，加劇溫室效應，造成全球暖化。

目前海平面升高是不可逆的，只能盡量減少溫室氣體產生來減緩升高速度，若能使用過去的數據去預測未來的海平面升高高度，可以先預防性的改變資源的利用、改善生存環境、計畫性遷都，提早作防範，減少海平面帶來的衝擊，獲取更多時間讓人類生存更長久。在海平面預測領域中，尚未發現有使用此研究數據去完成預測，但此研究數據對於海平面的變化都有密不可分的關係，都是必須要使用的數據，如：溫室氣體、溫度、南北極融冰。另外，在此研究數據中，已發現資料集的資料量不足，可能使現有的預測模型無法達到高準確率，所以本研究預期使用混和式預測模型，使其準確率可達到50%甚至更高的準確率，找出可使用在資料量不足的混和模型。

2 Related Literature

2.1 線性回歸

線性回歸應用於數據點中找到一條線，此線到所有數據點都是最短距離，通常使用在數據的趨勢或預測任務上。而在統計上則是使用最小平方法找多個自變數 (independent variable) 和一個應變數 (dependent variable) 關係建模的一種迴歸分析。只有一個自變數和一個應變數的情形稱為簡單線性回歸 (Simple linear regression)，大於一個自變數的情形稱為多元回歸 (multiple regression)。理論上自變數 (independent variable) 是不被其他變數影響的，只會去影響別人，所以被認為是「因」(Cause)。應變數 (dependent variable) 基本上是被其他變數影響的，被認為是「果」(effect)。簡單線性回歸在數據點中，得到的結果(應變數 y) 與來源變數 (自變數 x)可以用直線關係描述，如 (1) 所示。

$$y = \beta_0 + \beta_1 x \qquad (1)$$

其中 β_0 表示截距 (Intercept)而 β_1 表示斜率 (Slope)。我們可以利用統計方法中的最小平方法 (Least Square) 來找參數 β_0 和β_1，如 (2) 所示。最小平方法就是希望誤差的平方越小越好，取平方後皆為正值，所以最終期望所有訓練樣本的誤差平方和 (Sum Square error, SSE) 接近 0。

$$Loss(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \qquad (2)$$

2.2 長短期記憶與門控循環單元模型

長短期記憶 (LSTM) 是一種特殊的循環神經網路 (RNN)。與傳統的 RNN 不同，LSTM 使用三個不同的閥來控制單元的狀態，分別是輸入閥、遺忘閥和輸出閥。這三個閥在圖 2 中分別三個綠框表示。

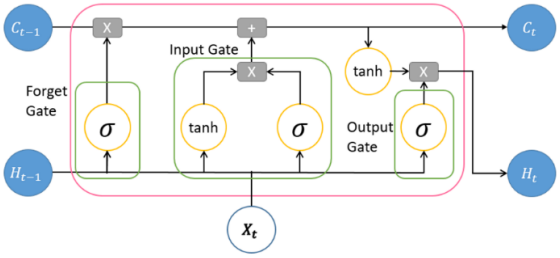

圖 1: LSTM 模型的示意圖。

遺忘閥通過 (3) 來控制遺忘，其中W_f和U_i代表要與前一個時間點的輸出和當前輸入相乘的權重矩陣，h_{t-1}代表前一個時間點的輸出，X_t代表當前輸入，b_f代表偏移量向量，所得的f_t可以決定哪些訊息應該被遺忘。輸入閥分為兩小部分，一部分稱為候選狀態向量$\tilde{\boldsymbol{c}}_t$和輸入閥向量$\boldsymbol{i}_t$，操作方法為 (4) 和 (5)，其中W_c, W_i, U_c 和 U_i代表權重矩陣，$\boldsymbol{b}_c$和 $\boldsymbol{b}_i$代表偏移量向量。

用這兩個向量$\tilde{\boldsymbol{c}}_t$和 $\boldsymbol{i}_t$來控制多少個單元狀態受到當前輸入的影響，新的單元狀態$\boldsymbol{c}_t$將由f_t, $\boldsymbol{c}_{t-1}$, $\boldsymbol{i}_t$和 $\tilde{\boldsymbol{c}}_t$決定，如 (6) 所示。輸出閥是為了控制將輸出多少個單元狀態，如(7)所示，這也是由當前的輸入$\boldsymbol{X}_t$和前一輪的輸出h_{t-1}決定的。最後，本輪的輸出向量h_t取決於本輪的單元狀態$\boldsymbol{c}_t$和輸出閥的向量$\boldsymbol{o}_t$，如(6)所示。由於這些閥的機制，LSTM 可以記住長期的依賴關係。大多數LSTM的輸出會是一個或多個向量，與地面實況相比，得到兩者之間的誤差，然後通過隨機梯度下降或其他優化算法矩陣更新網路中的權重。由於網路中存在多個閥，大大降低了部分分化過程中梯度消失或爆炸的可能性，這是 LSTM 比一般 RNN 的優勢。

$$f_t = \sigma(W_f h_{t-1} + U_i X_t + \boldsymbol{b_f}) \quad (3)$$
$$\tilde{\boldsymbol{c}}_t = tanh(W_c h_{t-1} + U_c X_t + \boldsymbol{b_c}) \quad (4)$$
$$\boldsymbol{i_t} = \sigma(W_i h_{t-1} + U_i X_t + \boldsymbol{b_i}) \quad (5)$$
$$\boldsymbol{c_t} = f_t \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \tilde{\boldsymbol{c}}_t \quad (6)$$
$$\boldsymbol{o_t} = \sigma(W_o \boldsymbol{h_{t-1}} + U_o \boldsymbol{X_t} + \boldsymbol{b_o}) \quad (7)$$

在 2014 年，Cho 等人為了改善 LSTM 執行速度較慢的問題，提出門控循環單元 (Gate Recurrent Unit, GRU) (Cho et al., 2014)，並且證明 GRU 可以加快模型執行速度與減少記憶體的使用。本研究將使用 GRU 模型進行後續實驗。

3 Dataset Collection and Processing

3.1 資料集說明

本研究使用 6 種資料集，包含：全球二氧化碳、全球氣溫、北極海冰層、南極海冰層、格陵蘭冰層和海平面高度，資料數據來源是由 NASA、satellite analysis at the University of Alabama、National Snow and Ice Data Center 和 NOAA 提供，對於各個資料集，首先使用線性回歸(Linear regression) 證明每一個資料集得數據是每年遞增或遞減的成長，了解其數據與時間的關係並證明資料都有時間性後，再使用 LSTM 時間序列模型做兩種實驗，第一是使用全球二氧化碳預測全球氣溫、全球氣溫預測北極海冰、全球氣溫預測南極與格陵蘭冰層、北極海冰層和南極海冰層預測海平面高度為第一研究數據。第二種是將全球二氧化碳、全球氣溫、北極海冰層、南極海冰層與格陵蘭冰層的所有數據，集結成一個大的資料集，再去做預測海平面高度的第二研究數據，最後做預測 30 年或 50 年的海平面高度。

資料名稱	蒐集時間	筆數
全球二氧化碳濃度	1958/3~2021/12	766 筆
全球平均氣溫	1978/12~2021/12	517 筆
北極海冰體積	1978/11~2021/12	518 筆
格陵蘭冰層體積	2002/1~2021/12	240 筆
南極冰層體積	2002/1~2021/12	240 筆
全球海平面上升高度	1993/1~2021/12	348 筆

表 1: 資料集數據

3.2 資料集處理

在全球二氧化碳、全球氣溫、北極海冰層、南極海冰層、格陵蘭冰層和海平面高度資料集中，每一份資料集都含有時間、目標資料與額外原研究的產出資料。在實驗中只需存留時間與目標資料，其餘欄位都先刪除；另外，在資料集南極冰層和格陵蘭冰層的年份與月份時間並無清楚標示，只標名年份與觀察順序，因此需要把觀察順序轉換成月份順序。

在資料集中，海平面資料是由四顆衛星所測得，從第一顆衛星收集數據到結束任務後，會由第二顆衛星繼續任務收集數據到結束任務，再由第三顆衛星接續任務，以此持續類推到第四顆衛星。目前第四顆衛星還在持續收集數據，每一顆衛星在每一年當中收集資料的次數皆不同，並且在每一顆衛星即將結束任務前，接續的衛星就會開始收集數據，因此資料集的每年資料量不同。於是先觀察第一顆衛星獨自接收的數據資料皆大於 31 筆，依照每年大月收集 3 次數據與小月收集 2 次數據的次數做計算，找出每一年份大於 31 筆的年份資料，並對當年的資料刪除到 31 筆。其刪除方式是測得當年資料是大於 31 筆，每一次刪除資料就比對是否剩下 31 筆，比對次數為基數次是刪除最大的數值，偶數次是刪除最小得數值，直到當年資料為 31 才結束並開始整理下一年資料。每年資料都整理為 31 筆後，再依照大月收集 3 次數據與小月收集 2 次數據的次數做成每一個月份的數據，得到最後的統一格式。最後把所有資料集合成一個大的資料集後，資料量是 240 組資料，每組資料有 6 個變項。

4 Experimental Results and Discussion

首先是使用簡單線性回歸的趨勢線來觀察資料的成長，圖 2 至圖 5 僅顯示全球二氧化碳、全球氣溫、北極海冰層面積、南極海冰層面積資料趨勢線，綠色是資料點，而紅色是趨勢線。從圖中，我們可以觀察到二氧化碳濃度與氣溫數據是逐年快速遞增，而冰層面積則是逐年快速遞減。

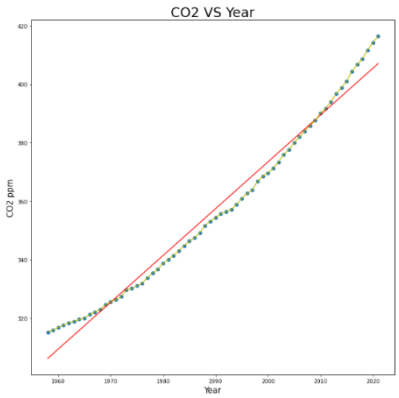

圖 2: 全球二氧化碳濃度。

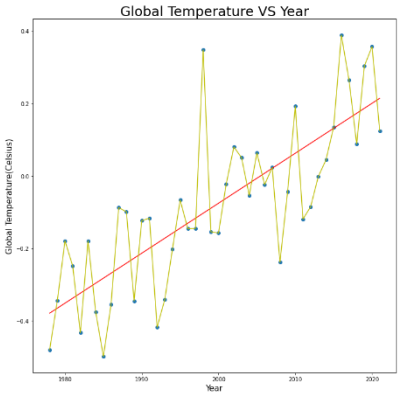

圖 3: 全球氣溫。

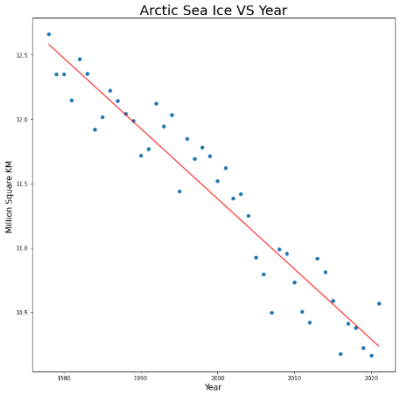

圖 4: 北極海冰層面積。

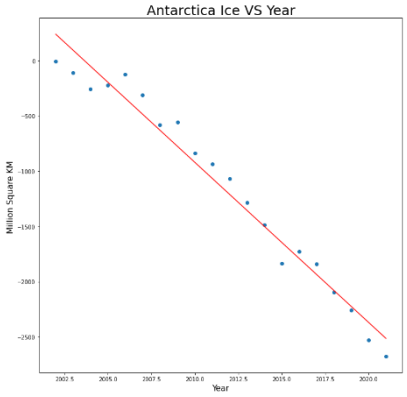

圖 5: 南極海冰層面積。

此外，我們評估 GRU 和 LSTM 模型，不同參數所取得之實驗結果。其中以 LSTM 3 層+MLP 2 層+Batch Size 32 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 2.11；以 GRU 2 層+MLP 2 層+Batch Size 64 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 **1.16**；以 GRU 2 層+MLP 2 層+Batch Size 32 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 1.17；；以 GRU 2 層+MLP 2 層+Self Attention 層+Batch Size 32 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 1.22；以 GRU 3 層+MLP 2 層+Self Attention 層+Batch Size 32 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 1.21。

5 Conclusion and future work

在這項研究中，我們使用 6 種資料集，包含：全球二氧化碳、全球氣溫、北極海冰層、南極海冰層、格陵蘭冰層和海平面高度用以建立海平面高度預測模型。其結果可進一步用於環境監測，為人們提供環境保護相關因子建議，用以降低汙染環境之因子。最後，實驗結果表明，以 GRU 2 層+MLP 2 層+Batch Size 64 所取得的平均絕對百分比誤差 (MAPE, Mean Absolute Percentage Error) 為 **1.16**，取得了最好的系統性能。

在未來的工作中，我們希望能獲得更多的環境因子收集數據集，以便我們能訓練出更適合之海平面高地預測系統。

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. *Yara parser: A fast and accurate depen-dency parser*. *Computing Research Repository,* arXiv:1503.06733. Version 2.

Hui-Hsin Tseng, Chao-Lin Liu, Zhao-Ming Gao, and Keh-Jiann Chen. 2002. 以構詞律與相似法為本的中文動詞自動分類研究 (a hybrid approach for automatic classification of Chinese unknown verbs) [in Chinese]. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 7, Number 1, February 2002: Special Issue on HowNet and Its Applications*, pages 1–28.