AI CUP 2024 秋季賽
根據區域微氣候資料預測發電量競賽報告

隊伍：TEAM_5883
隊員：陳昱翔、藍中崑、陳偉俊、陳伃汶（隊長）
Private leaderboard：645445.1 / Rank 25

壹、環境

作業系統：Ubuntu Linux V24.04.1 LTS
CPU：13th Gen Intel(R) Core(TM) i7-13700K
GPU：NVIDIA GeForce RTX 4090
CUDA：12.6
語言：Python 3.12.2
套件：

- numpy v1.26.4
- pandas v2.2.2
- matplotlib v3.9.2
- seaborn v0.13.2
- scikit-learn v1.5.1
- pytorch v2.5.1
- joblib v1.4.2

額外資料集：CODiS 氣候觀測資料查詢服務網站之氣象資料

貳、演算方法與模型架構

此部分為此模型的詳細設計說明，內容包含：「整體架構設計理念」、「核心架構解析」、「關鍵參數設定」。

一、整體架構設計理念

依據比賽設計，參賽者得使用先前給予的資料集，訓練一套估算模型，再根據題目進行過往太陽能發電之電量預測。

首先，由於題目只給時間及測站資訊，並且考慮比賽目的為估算已發生之現有數據，而非預測未來，故我們決定使用一般回歸模型(classic regression model)，而非範例程式碼的 LSTM。

接著，在選擇回歸模型的過程中，我們考慮了 Random Forest, XGBoost, LightGBM, CatBoost...等方法，經過討論及嘗試後，最終決定使用 Feed-Forward Neural Network (FNN) 。

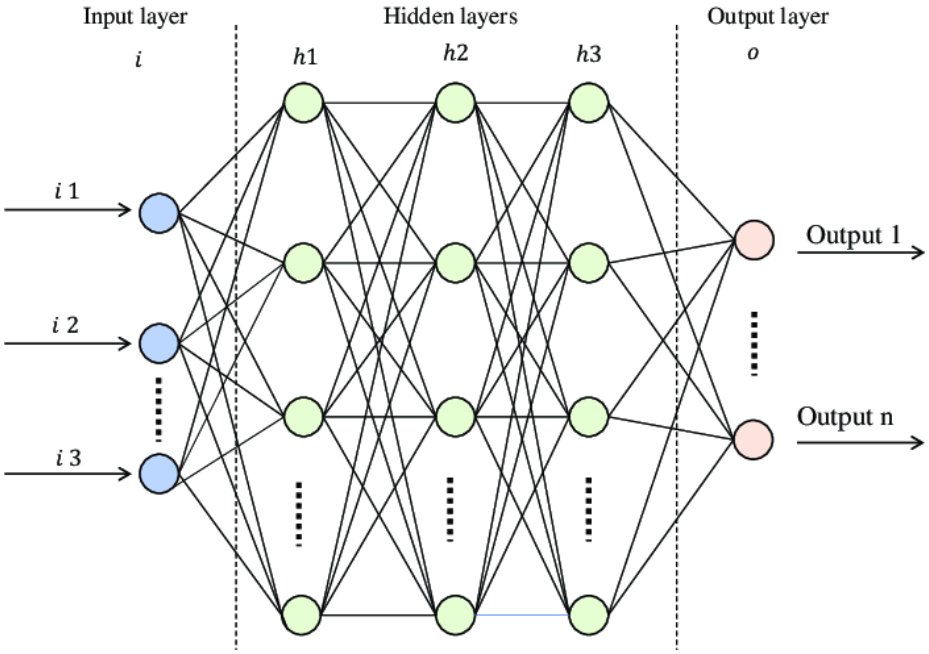

圖 1 FNN 模型架構

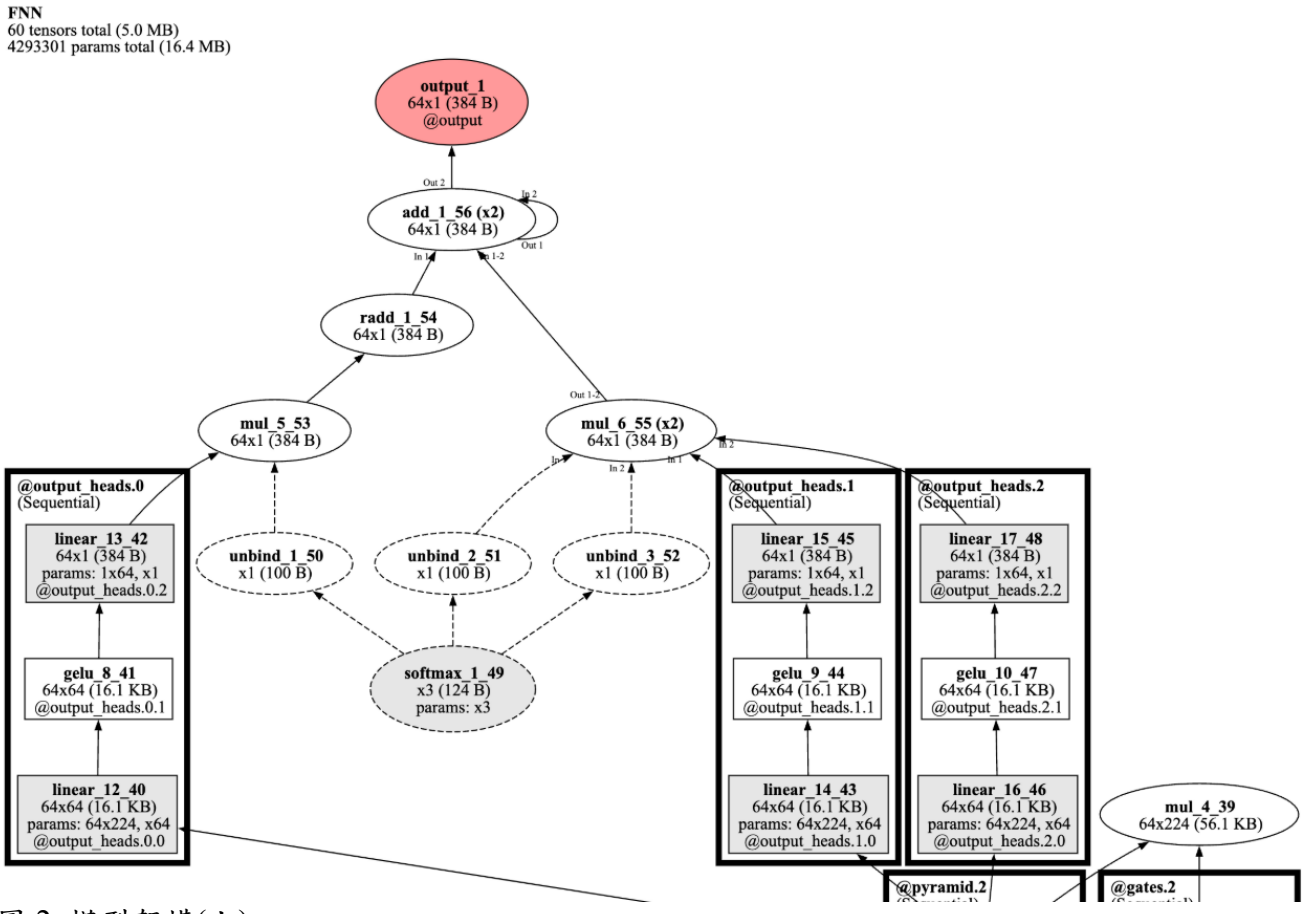

圖 2 模型架構(上)

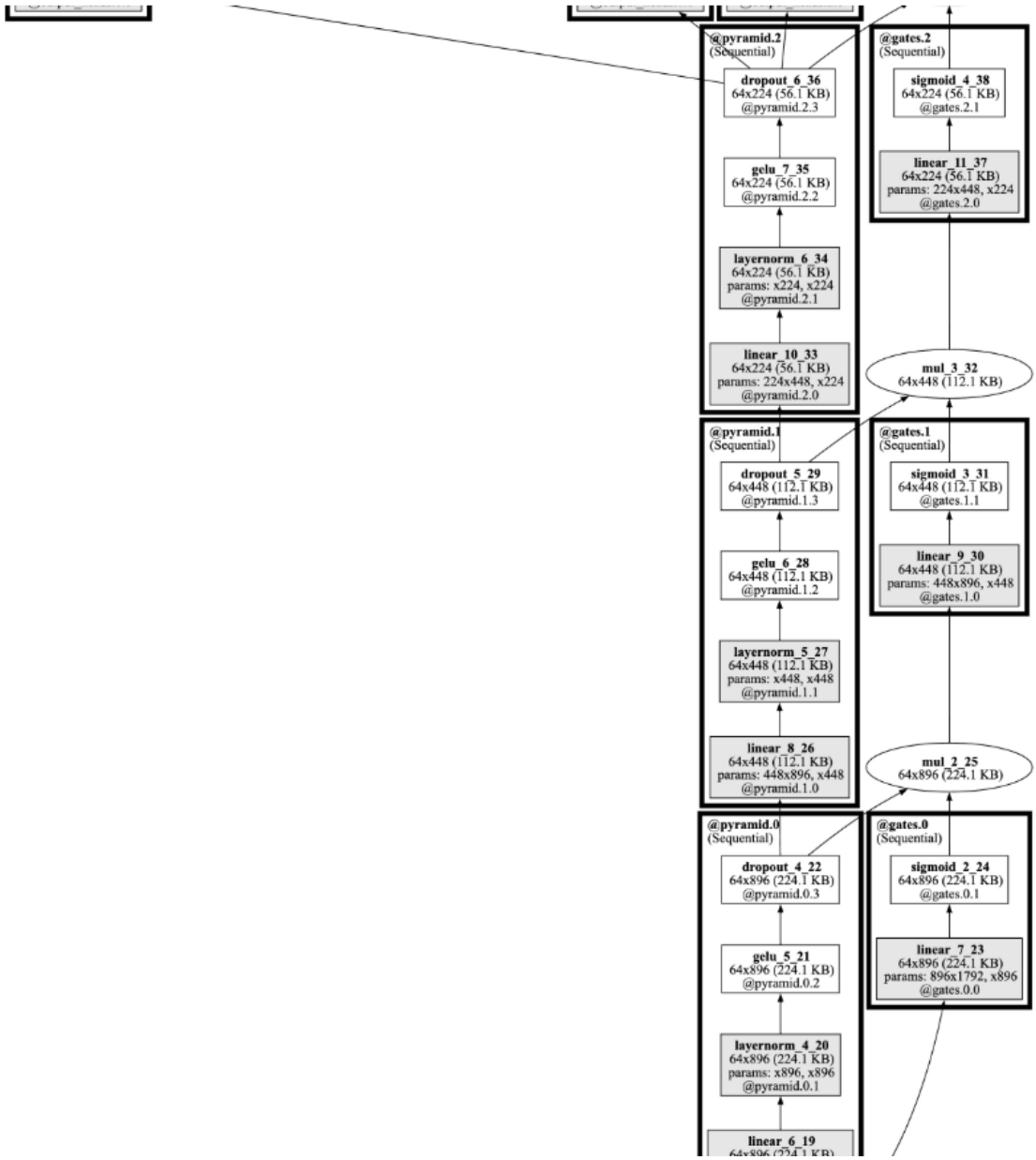
@pyramid.2
(Sequential)
dropout_6_36
64x224 (56.1 KB)
@pyramid.2.3
gelu_7_35
64x224 (56.1 KB)
@pyramid.2.2
layernorm_6_34
64x224 (56.1 KB)
params: x224, x224
@pyramid.2.1
linear_10_33
64x224 (56.1 KB)
params: 224x448, x224
@pyramid.2.0
@gates.2
(Sequential)
sigmoid_4_38
64x224 (56.1 KB)
@gates.2.1
linear_11_37
64x224 (56.1 KB)
params: 224x448, x224
@gates.2.0
mul_3_32
64x448 (112.1 KB)
@pyramid.1
(Sequential)
dropout_5_29
64x448 (112.1 KB)
@pyramid.1.3
gelu_6_28
64x448 (112.1 KB)
@pyramid.1.2
layernorm_5_27
64x448 (112.1 KB)
params: x448, x448
@pyramid.1.1
linear_8_26
64x448 (112.1 KB)
params: 448x896, x448
@pyramid.1.0
@gates.1
(Sequential)
sigmoid_3_31
64x448 (112.1 KB)
@gates.1.1
linear_9_30
64x448 (112.1 KB)
params: 448x896, x448
@gates.1.0
mul_2_25
64x896 (224.1 KB)
@pyramid.0
(Sequential)
dropout_4_22
64x896 (224.1 KB)
@pyramid.0.3
gelu_5_21
64x896 (224.1 KB)
@pyramid.0.2
layernorm_4_20
64x896 (224.1 KB)
params: x896, x896
@pyramid.0.1
linear_6_19
@gates.0
(Sequential)
sigmoid_2_24
64x896 (224.1 KB)
@gates.0.1
linear_7_23
64x896 (224.1 KB)
params: 896x1792, x896
@gates.0.0

圖 2 模型架構(中)

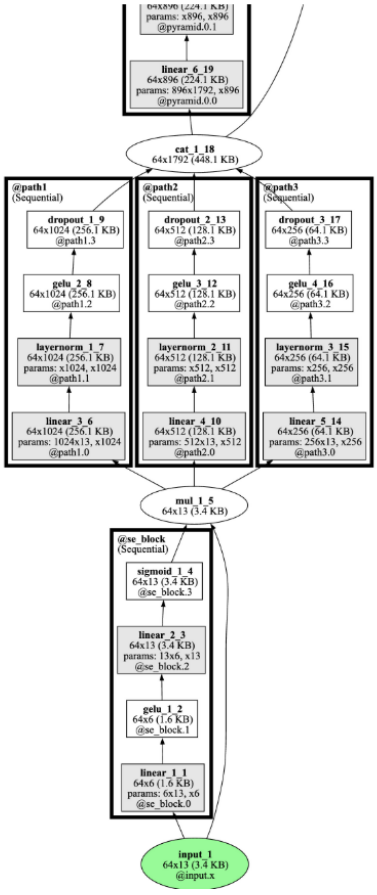

圖 2 模型架構(下)

二、核心架構解析

此部分將分為「資料預處理層」、「多重特徵提取系統」、「金字塔式資訊提煉結構」、「閘門結構」、「多節點整合預測機制」，五個部分說明。

1. 資料預處理層：

   我們使用了「SENets」（Squeeze-and-Excitation）進行預處理，此 Network 像是個智慧過濾器，可以識別並強調重要特徵，同時降低無關特徵的影響。具體而言，其運作方式為：先將資料壓縮至原先的一半大小，進行分析，最後再還原回原始維度。

2. 多重特徵提取系統：

   進行資料處理時，我們設計了三條並行的資料處理通道，分別為：

   - 深層通道（1024 個節點）：捕捉複雜的整體特徵
   - 中層通道（512 個節點）：處理中等規模的特徵模式
   - 淺層通道（256 個節點）：關注局部細節特徵

   此外，每個通道都配備有「特徵正規化層」、「激活函數（GELU）」、「動態降噪機制（Dropout）」。特徵正規化層，用以確保資料分布的穩定性；激活函數，可以提供非線性的轉換能力、動態降噪機制，用以預防過度擬合。

3. 金字塔式資訊提煉結構：
   資料處理過程中，關於資料提煉的部分，我們採用了金字塔式的資料提煉結構，總共三層：
   第一層：1792 → 896 個節點
   第二層：896 → 448 個節點
   第三層：448 → 224 個節點
   此外，每一層都包含：
   「資料轉換層」、「正規化處理」、「非線性(GELU)激活」、「動態降噪機制（Dropout）」。在此部分，動態降噪機制存在三層，其參數分別為：0.25, 0.2, 0.15。

4. 閘門結構：
   閘門結構用以對應 pyramid layers，可調節信息流通量，每一層都包含 Sigmoid，並且可分為三層：
   - 第一層閘門：1792 → 896 個節點
   - 第二層閘門：896 → 448 個節點
   - 第三層閘門：448 → 224 個節點

5. 多節點整合預測機制：
   此機制存在三個獨立預測單元（output_head），此外，每一單元大小依序為: 224 → 64 → 1

三、關鍵參數設定

此部分將分為「學習引擎設定」、「訓練過程控制」、「特殊處理機制」，三個部分說明。

1. 學習引擎設定：
   我們採用了 AdamW 優化器，將基礎學習率設為 0.001，權重衰退設為 1e-5。並使用了動態學習率調整系統 OneCycleLR，採取週期性調整策略，將預熱階段設為 30 %，epoch 設為 100
2. 訓練過程控制：
   其中，批次處理機制中將 batch 的大小設為 64，Loss Function 採用 MSE Loss，訓練時 epochs 訂為 100。進行梯度裁剪，最大範數 max_norm 設為 1.0；使用早停機制：觀察期設為 20 輪，並將判定標準設為驗證集的損失值。
3. 特殊處理機制：
   將特徵正規化，以確保各個特徵的尺度具備一致性；控制梯度，為防止在訓練過程中，發生梯度爆炸；採用 Kaiming 初始化方法，將權重初始化。

參、創新性

從「貳、演算方法與模型架構」中，具創新設計的特點有：

1. 資料處理的多樣性：多重特徵提取、多層次資訊整合
2. 穩定性保障機制：多重正規化層、動態降噪系統、梯度調節機制

3. 靈活的自適應能力：特徵重要性自動調節、學習率動態調整、預測權重自適應分配

此神經網絡模型，在以下列舉的面向具創新性：

1. 多重視角的特徵學習：
   模仿人類在觀察事物時會同時注意到大局和細節，此模型設計了三條平行的學習通道。好比有三個專家：一個觀察整體格局（1024 個觀察點），一個關注中等細節（512 個觀察點），一個專注最細微的特徵（256 個觀察點）。這種設計讓模型能夠同時從不同層面理解數據，猶如全方位地理解問題。

2. 特徵重要性的自適應調節：
   模型中的「擠壓—激勵機制（Squeeze-and-Excitation）」有如智能過濾器。它能自動判斷哪些特徵在當前預測中相對重要，哪些較不重要，並相對應地調整它們的影響力。這就像是位經驗豐富的專家，知道在不同情況下應該關注什麼重點。

3. 金字塔式的知識提煉：
   此設計的靈感來自於人類思維中的抽象化過程：當我們在理解複雜概念時，會逐步將信息提煉成更簡潔但更有意義的形式。模型通過逐層減少特徵數量（從 1792 到 896，896 到 448，448 到 224），實現了信息的逐步提純，僅保留最有價值的資訊。

4. 多輸出節點預測機制：
   好比請了三位專家同時對同一問題進行預測，然後根據每位專家的可信度來綜合他們的意見。模型會自動學習如何最佳地結合這三個預測結果，這種機制提高了預測的穩定性和準確性。

5. 自適應學習策略：
   此模型採用了靈活的學習速率調整機制，效法因材施教：學習初期可以大膽嘗試，隨後逐漸變得謹慎。此策略能在保證學習效果的同時，避免過度擾動已經學到的知識。

關於特徵相關設計中，具創新性的部分有：

1. 原先，依照 AICUP 提供的可視化和描述，明顯大多數功能都相當不可靠（例如：感測器相關問題、光照量具有採集上限等）。此外，在「upload.csv」中沒有為我們提供這些用以訓練的特徵。
   因此，我們決定利用外部天氣信息進行預測。而不是先依「upload.csv」中的時間，冒險預測其對應的氣象特徵，查詢這些不可靠的特徵後，再預測對應發電量。

2. 從視覺化中，我們可以看到與陽光相關的特徵似乎對發電影響最大。這促使我們使用來自 *[4]* 「CODIS（氣候觀測資料查詢服務）」的數據，其提供一年內每日每時的：溫度、日照時長，甚至輻射資訊，並確定包含了訓練資料和提交資料「upload.csv」中的時間區間。此外，此網站提供的數據，還能填

補測站 17，因是私人民宅而未知的資訊：因 17 號測站距離這個氣象站僅 2.6 公里。

圖 3 交通部中央氣象署花蓮氣象站

3. 關於資料特徵擴充，我們認為具有創新性的部分為：
   - 利用正弦函數 sin 和餘弦函數 cos，調整小時、月份等具週期性的時間特徵，以及感測器方向。避免數值中最小值和最大值被視為差距很大的幻覺，以感測器方向為例：實際上 0 度和 360 度為同一方向，但由於數值差，會誤以為向距甚遠。
   - 因題目提供的測量時間有跨度不同季節，故利用月份計算季節
   - 位置 17 我們猜測其高度為三樓，此為各測站多為 1 樓、5 樓，故取其平均作為設定

肆、資料處理

此部分為使用的資料處理方式及理念說明，內容包含：「資料載入與整合策略」、「資料清理修正」、「特徵工程與增強」、「特殊處理機制」。

一、資料載入與整合策略

1. 資料整合機制

   在基礎資料載入的部分，我們先處理了比賽方所提供的，多個測站（L1—L17）的資料，並設計成可支援多種檔案格式（如：一般檔案和 _2 檔案）。

2. 外部資料引入

   我們將原有模型，整合氣象局 CODIS 系統的資料：各月份中，每日、每時之資料：

   - 溫度資料
   - 日照時間
   - 輻射量測資料

二、資料清理與修正

1. 異常值處理

   將以下資料，針對個別的合理範圍作篩選：

   - 氣壓：950-1030 hPa
   - 溫度：0～45°C
   - 濕度：10-100 %

2. 缺失值處理策略

   分為以下兩類資料作處理：

氣象資料補充：
使用平均值填補缺少的資料(如：以小時資料，填入每分鐘)
位置相關資料：
因為提供 17 號測站的高度位置，考慮民宅普遍高度，加上各測站為 1 樓或 5 樓，故取其平均，猜測 17 號測站高度為 3 樓。

三、特徵工程與增強

1. 時間特徵進行擴展

基礎時間特徵：

- 年、月、日
- 小時、分鐘
- 星期幾

2. 以月分計算季節性特徵

季節依以下月份進行劃分：

- 冬季：12-2 月
- 春季：3-5 月
- 夏季：6-8 月
- 秋季：9-11 月

3. 週期性特徵轉換

以下為具此特性的循環時間：

- 小時週期（24 小時制）
- 月份週期（12 個月）

加以使用正弦函數和餘弦函數轉換：

- hour_sin = $\sin(2\pi \times (\text{hour} + \text{minute}/60)/24)$
- hour_cos = $\cos(2\pi \times (\text{hour} + \text{minute}/60)/24)$
- month_sin = $\sin(2\pi \times \text{month}/12)$
- month_cos = $\cos(2\pi \times \text{month}/12)$

4. 位置特徵增強

方向編碼：將各測站位置的方向角度轉換為正弦/餘弦值，以處理方向所具有的循環性質。
高度分類：
將建築物高度分類為：1 樓、3 樓、5 樓

四、特殊處理機制

1. 日照飽和處理

設定最大閾值為：117758.2，超出或包括此數值的狀，標記其為飽和狀態

3. 資料品質保證

設置其具有自動化的異常檢測、處理過的詳細統計資訊之輸出、對各個資料的常理範圍進行驗證。

伍、訓練方式

在此段落，我們將分為：「訓練準備階段」、「核心訓練策略」、「訓練過程控制」、「進階訓練技巧」、「訓練監控與評估」、「訓練後處理」，此六個部份分別討論。

一、訓練準備階段

首先，關於資料集切分策略：

我們採用標準隨機切分，將訓練集合測試集依 8：2 的比例作切割。使用固定隨機種子：42，以確保模型的可重現性存在。並在訓練模型時進行隨機打亂：shuffle＝True

接著，關於批次處理設定：

我們使用使用 DataLoader 進行批次加載，將訓練集啟用隨機打亂，並使測試集保持順序：shuffle＝False。

以上方法雖概念簡單、易實現，但由於此模型不是用以預測未來值，而是在已知條件下預測發電量，並且各個時間頂的數據相對獨立，故我們認為可應用在此模型上，並取得不錯的結果。

最後，關於特徵標準化處理：數值特徵方面，將其標準化至零均值、單位方差；週期性特徵方面，則保持三角函數轉換後的尺度。

二、核心訓練策略

1. 學習率管理採用 OneCycleLR 策略：最大學習率設為 0.001，預熱階段設為總周期的 30%，降溫階段採取漸進式降低。
2. 優化器配置使用 AdamW 優化器：基礎學習率設為 0.001，權重衰減設為 1e-5，梯度裁剪之最大範數設為 1.0。
3. 損失函數設計方面，我們主要採用 MSE 損失，以考慮預測值的規模和分布

三、訓練過程控制

1. 早停機制之觀察指標為驗證集損失，耐心值設為 20 個周期，儲存策略為保存最佳模型。
2. 優學習過程監控的部分有：即時損失追蹤、訓練集與驗證集性能對比、學習曲線分析。
3. 模型檢查點可以定期保存模型狀態、作為最佳模型之保存機制、訓練狀態恢復功能。

四、進階訓練技巧

1. 梯度管理有梯度裁剪的實作
2. 正則化策策略有權重衰減、Layer Normalization、學習率動態調整、及使用多層 Dropout：

- 深層路徑：0.3
- 中層路徑：0.2
- 淺層路徑：0.1

五、訓練監控與評估

1. 性能指標追蹤使用了以下幾種指標：MSE（均方誤差）、RMSE（均方根誤差）、MAE（平均絕對誤差）、R^2分數
2. 訓練過程的視覺化應用在損失變化的追蹤、學習率變化的追蹤
3. 模型診斷主要有是否存在過擬合/欠擬合的檢測、特徵重要性分析、預測誤差分析

六、訓練後處理

1. 模型評估有針對完整測試集評估、跨位置性能分析、時間序列預測能力評估
2. 結果分析有進行預測誤差模式分析、特徵影響力分析、模型穩定性評估
3. 部署準備的部分作了模型的壓縮與優化、提升推理效率、部署環境適配

陸、分析與結論

我們將依循模型設計的過程及修正，一步步詳細解釋其演變與緣由：

一、17 個獨立小模型

首先，將比賽方提供的 17 個測站，先分別將各個測站，訓練出各別測站獨立的小模型，此時總誤差約 150 萬。但此時我們發現第一個問題：比賽方公告的題目中，並不包含微氣候資訊。因此，我們嘗試了一些方法來填補缺漏資料：

1. 使用同一測站目標時間的前後幾個小時
2. 若無，找同方向、同樓層的鄰近測站
3. 若仍舊無法解決，使用全部有資料的測站，同一時間的資料作平均

使用上述方式後，仍存在問題：17 個測站各自資料筆數相差頗大，此外，即便估算每測站平均八萬筆資料，資料筆數依舊略少，促使此時的模型存在，因數據數量不足而產生的誤差。

二、用一個大模型並簡化特徵

接著，我們嘗試著將 17 個測站資料，用於同一模型做訓練，然後嘗試純粹使用時間及測站地點資訊。本次嘗試是因為我們認為，題目給予的時間數值，本就隱含不同時段（例如：月份、日期、小時）的氣象資訊，故希望先藉由簡化特徵類型，確認各類資訊的可能影響性。模型方面，我們選擇使用 FNN 回歸模型，因為此模型學習能力最好、最全面，什麼類型的應用都可以學習。

綜合上述的兩大變動，此時的結果比原先切成 17 個獨立模型的效果要優秀，此時總誤差約 150 萬，因為此時只使用了時間資訊，進行發電量預測，即能得到和原先的模型差不多的效果。爾後，我們往這方面繼續改進，例如：加入「季節」、「樓層」、「感測器方向（經過三角函數轉換後）」等新特徵，效果也如預期那般令人滿意，此時總誤差進步到 110 萬。

接著，分析特徵彼此相關性後，發現日照量對模型準確度影響十分明顯。同時，由於考量到氣溫也會間接反映日照的變化，所以我們試圖加入日照、氣溫等更多特徵，但由於比賽方給予的資料集，有不少缺漏，且分布不均，故我們遇到了新的問題：該如何使用不完整且難以填補的資料。

三、加入重要 API，取得重要改進

在決定使用日照、氣溫後，我們先初步嘗試使用比賽方提供的微氣候資料，但發現測站設定的日照量最大上限經常達標；加上指定預測題目中的序號，只有提供時間和測站的資訊。若要使用日照等相關資訊，必須先預測出缺少的微氣候資訊，再用以預測發電量。

因此，為了避免加劇幻覺，我們在搜尋相關氣象資源時，發現「CODIS」此氣象網站，此網站可提供整年、整月、逐日、每小時的資料。加上此網站的設

站位置，在花蓮市區內，剛好距離 17 號測站僅 2.6 公里，可以完整補齊 17 號測站因其為民宅資訊不公開的部分，故這個 API 非常理想。

其中，我們選用「氣溫」、「每小時日照時常長」、「輻射量」，這三類資訊，作為新特徵使用。加入後，模型結果大幅改進其效能，此時總誤差減少至 80 萬。

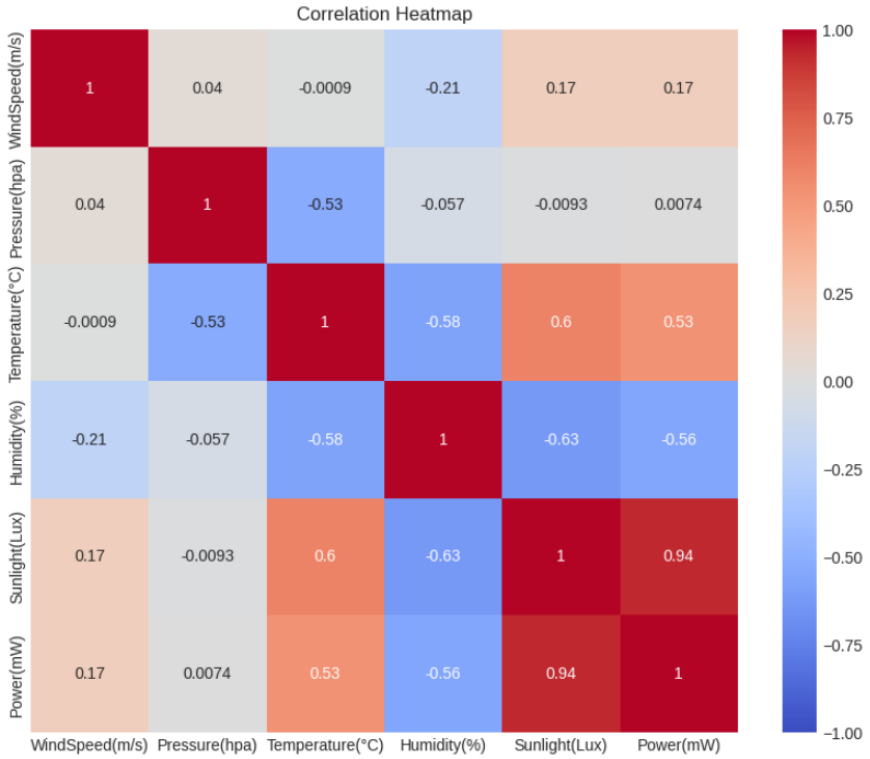

圖 4 相關係數分析圖，結果可見日照和發電量高度相關

四、參數優化及架構調整

最後，我們對這個模型持續進行優化，例如：調整節點數量。在此過程中，我們發現訓練過程因參數、節點數量增加，導致訓練困難。為了解決此問題，所以我們加入學習率動態調整機制（one cycleLR ）試圖解決。

但加入並訓練後，我們觀察到，雖然 training lost 下降，但是 validation lost 卻沒隨之下降，也發現存在過擬合的問題。因此，再加入 Drop out layer，希望能避免過擬合的發生，加入後也確實實現。

在解決過擬合問題後，接著，我們參考了 LSTM、RNN 等處理時間序列資料的模型，決定嘗試加入其特色：閘門機制，來處理時間序列性質的資訊。加入後，我們的模型也確實有所精進。

最後，我們使用了生成式 AI （ ChatGPT），與其討論還有哪些可改善空間。因此，我們參考了其建議，加入多重通道機制：得以更好地提取淺層、中層、深層的資訊。

在做了以上一系列的優化後，此時的總誤差來到目前最佳結果：54 萬。

圖 5 與生成式 AI 討論還能如何增進效能

五、嘗試 transformer

後續我們曾嘗試使用 transformer，因為此架構使用注意力機制，考慮其可能可以更好地考慮到，和目標特徵相鄰的資訊，故做此嘗試。但實現後，成績也只有 60 萬，未能超越原先架構的成果：54 萬。

可惜的是，由於比賽時間限制，未能做後續優化。考慮到尚未進行優化，transformer 的效果就已接近原先的模型結果。因此，我們預期未來若針對 transformer 模型，或將參數優化，效果理應可超越原來架構的結果，

六、結論

在整個模型設計與優化過程中，我們逐步改進了數據處理和架構設計。最初，依據比賽的要求，使用 FNN 作為基礎模型，因其能靈活捕捉非線性關係，並適應我們的多元特徵結構。模型優化的關鍵點包括數據的特徵工程、參數調整及架構設計。

最大的突破是整合氣象局 API，它提供了精確的氣溫、日照時長和輻射資訊，顯著改善了模型的特徵完整性，使預測誤差從 110 萬下降到 80 萬。此外，我們採用了 OneCycleLR 動態學習率、Dropout 機制，解決了過擬合問題，並引入了生成式 AI 建議，構建多通道機制，最終將誤差降至 54 萬。

最後雖然嘗試的 Transformer 架構尚未超越現有模型，但也為未來精進改善提供了方向。

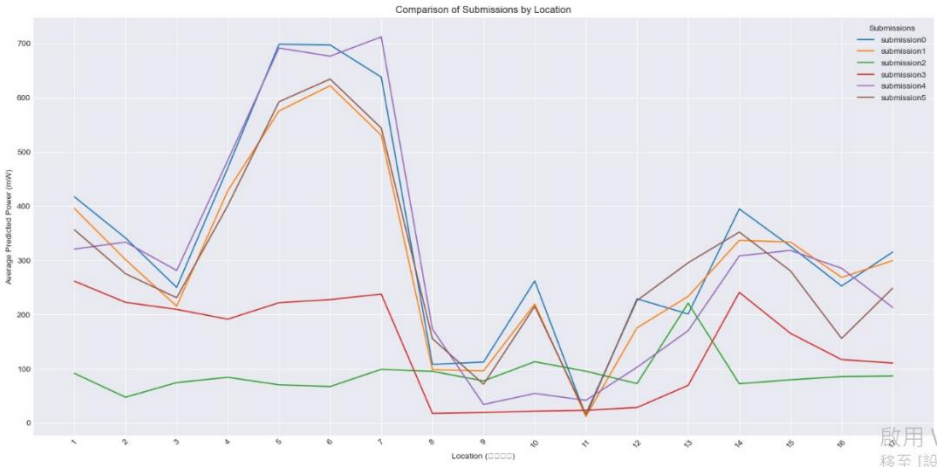

圖 6 逐次改進模型後的各測站平均預測結果

predictions1945_pyramid_100epoch.csv pyramid 100 epoch 上傳成員 陳 偉俊	2024-11-28 08:06:35	545343.08	645445.15	Scoring success.

圖 7 上傳的最好成績

柒、程式碼

程式碼資源與實驗重現教學於 Github 連結：
https://github.com/tanerijun/aicup-2024-fall

捌、使用的外部資源與參考文獻

[1] N.-T. Nguyen et al., "Solar Radiation Forecasting Based on Random Forest and XGBoost," 2024 7th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, Vietnam, 2024, pp. 136-140

[2] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[3]Hendrycks, Dan, and Kevin Gimpel. "Gaussian error linear units (gelus)." *arXiv preprint arXiv:1606.08415* (2016).

[4] CODiS 氣候觀測資料查詢服務. (n.d.). CODiS 氣候觀測資料查詢服務. https://codis.cwa.gov.tw/

[5]J. Gaboitaolelwe, A. M. Zungeru, A. Yahya, C. K. Lebekwe, D. N. Vinod and A. O. Salau, "Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison," in IEEE Access, vol. 11, pp. 40820-40845, 2023.

作者聯絡資料表

隊伍名稱	5883	Private Leaderboard 成績	645445.1	Private Leaderboard 名次	25
身分 （隊長/隊員）	姓名 (中英皆需填寫) (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號)	學校＋系所中文全稱 (請填寫完整全名，勿縮寫)	學校＋系所英文中文全稱 (請填寫完整全名，勿縮寫)	電話	E-mail
隊長	陳仔汶 Yu—Wen,Chen	國立臺灣師範大學資訊工程所	National Taiwan Normal University Department of Computer Science & Information Engineering	0912-971-290	61347093s@gapps.ntnu.edu.tw
隊員 1	陳偉俊 Vincent Taneri	國立臺灣師範大學資訊工程所	National Taiwan Normal University Department of Computer Science & Information Engineering	0981-823-282	tanerivince@gmail.com
隊員 2	藍中崑 Chung—Kun,Lan	國立臺灣師範大學資訊工程所	National Taiwan Normal University Department of Computer Science & Information Engineering	0900-398-978	chqueen.tw@Gmail.com
隊員 3	陳昱翔 -Yu-Hsiang,Chen	國立臺灣師範大學資訊工程所	National Taiwan Normal University Department of Computer Science & Information Engineering	0963-910-130	rain910130@gmail.com

★註 1：請確認上述資料與 AI CUP 報名系統中填寫之內容相同。自 2023 年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信 moe.ai.ncu@gmail.com。報告繳交截止時間後將不予修改。

★註 2：繳交程式碼檔案與報告，請 Email 至：ailabailab5051@gmail.com，並同時副本至：t_brain@trendmicro.com 與 moe.ai.ncu@gmail.com。缺一不可。