1. google 所提出的 BERT 就是一種自監督式 pre-train model，雖然可以有不同應用場景，但最常被應用在 NLP 模型上，特別是在訊號處理，語音辨識，或文字生成上，它所運用的技術主要是兩個層面上，一是做填空題，首先隨機的對一些 input 做 mask 處理，以文字為例便是隨機對個別某些字遮蔽(你也可以說是挖洞)，然後這些遮蔽字(token)透過 BERT 給定 embedding 與分類器做完線性變換後通過 softmax 輸出一個各個字分布機率的向量，再與 mask 前的 input 做 cross entropy 求最小值。另一則是預測一個句子是是另一個句子的下一句 next sentence prediction，將可能是兩個句子的文字組合，加入兩個 token SEP CLS 分別表示分隔與起點，最後透過 BERT 與分類器得出 YES 或 NO。

2. 部分同意，在走迷宮這類問題上，使用到相當多強化學習類型的模型，透過獎勵與懲罰機制來訓練 agent，舉例來說以 QLearing，模型內有一依據環境不斷更新的 Qtable 記錄了狀態 state 與動作 action 的對應表以及接下來各個決策預期所能得到的獎勵或懲罰，依照環境更新的刺激不斷地給予 agent 新的回饋，以此來訓練 agent 學習在當下的狀態做出何種動作的決策，這種類型的模型已經在走迷宮的研究行之有年，也多經常能解決設定好的迷宮，但我認為我們仍難以證明這種方式所找到的解必定會是最短路徑解，因此只能說部分同意。

3. 可能的，儘管 chatGPT 給了一個只能針對某些演算法或特定目標或特定資料最佳化，而且必須取決於問題的本質與演算法自身的回答。但我認為更精確地說只要是未被數學證明為最佳解的演算法，都仍有機會可以做進一步的最佳化，舉例來說圍棋作為人類最古老的遊戲，也能視為可以透過演算法處理的問題，惟其變化的可能性與計算量太高，人類有限的大腦必須透過固定的範式來對遊戲進行理解，這在圍棋術語中稱為定石，這些定石都是經過時間長河的棋士經驗與直覺累積形成的行之有年的思考模式代代傳授，我們數學知識體系與演算法也有異曲同工之妙，在那些相當依賴經驗與直覺的領域特別如此，充滿了人類思考模式的特質。然而 alphago 的誕生徹底的改寫了長久以來許多人類視之為圍棋常識的定石，正如同飛機不需要模仿鳥類振翅飛行，潛水艇不需要模仿魚類擺鰭游泳，人工智慧也不需要模仿人類思考，並且其上限也只受硬體與演算法限制，所以也許也很有機會能夠打破人類舊有的思考框架，拓展至更多人類思考的未竟之地。

4. $w_0 = T_+ - T_-$

pf.

Assume W_f^T is the solution, let $W_0^T = 0$, the zero-th component $x_0 = 1$

$\forall x_i$, $i = 1, 2, \cdots n$, $x_i \in D$, known $W_{t+1} \leftarrow W_t + y_{n(t)} x_{n(t)}$ where We only update when $y_{n(t)} W_t^T x_{n(t)} \le 0$, thus total number of "update" is $T_+ + T_-$

$W_f^T \leftarrow y_{f-1} x_{f-1} + W_{f-1}^T \mid \cdots \mid W_1^T \leftarrow y_1 x_1 + W_0^T$

$(T_+ + T_-)$ times

where component w_0 of W_f^T : $w_0 = \underbrace{y_{n(f-1)} x_n + \cdots}_{\#\text{ of } y_{n(t)} = 1 : T_+,\ \#\text{ of } y_{n(t)} = -1 : T_-} + 0$, where $y_{n(i)} x_{n(i)} = \begin{cases} -x_{n(i)} & \text{if } y_{n(i)} < 0 \\ x_{n(i)} & \text{if } y_{n(i)} \end{cases}$

since $x_{0i} = 1 \ \forall x_i \in D$, $w_0 = \underbrace{1+1+1\cdots}_{\#\text{ of } 1 : T_+} \underbrace{-(1+1+1\cdots)}_{\#\text{ of } -1 = T_-} = T_+ - T_-$ //

5

pf. Assume perfect solution is W_f^T, known $W_f^T W_{t+1} \ge W_f^T W_t + \min_n y_n W_f^T x_n$

$\ge W_f^T W_{t-1} + 2(\min_n y_n W_f^T x_n) \Rightarrow W_f^T W_T \ge W_f^T w_0 + (T)(\min_n y_n W_f^T x_n)$

Let $w_0 = 0$, $W_f^T W_T \ge T \cdot (\min_n y_n W_f^T x_n)$ where T is total # of "mistake"

And we know $\|W_{t+1}\|^2 \le \|W_t\|^2 + \max_n \|x_n\|^2 \Rightarrow \|W_T\| \le \|w_0\|^2 + T(\max_n \|x_n\|^2)$

Thus $\|W_t\| \le \sqrt{T}(\max_n \|x_n\|)$

$\therefore 1 \ge \dfrac{W_f^T W_T}{\|W_f\| \|W_T\|} \ge \dfrac{\min_n y_n W_f^T x_n \cdot T}{\|W_f\| \cdot \max_n \|x_n\| \cdot \sqrt{T}}$

$\therefore \dfrac{W_f^T}{\|W_f\|}, \dfrac{W_T}{\|W_T\|}$ are unit vector, $\dfrac{W_f^T W_T}{\|W_f\| \|W_T\|} \le 1$

$$\therefore 1 \geq \frac{T \cdot \min_n y_n W_f^T X_n}{\sqrt{T} \cdot \max_n \|X_n\| \cdot \|W_f\|}, \quad T \leq \frac{(\max_n \|X_n\| \cdot \|W_f\|)^2}{(\min_n y_n W_f^T X_n)^2}$$

Since $\max_n \|X_n\|$ is maximum Euclidean distance of all data from f there are m of d distinct words from X_n at most.

$X_n = \begin{bmatrix} x_0 \\ \vdots \\ \vdots \end{bmatrix}$ d-dimention, number of 1 : m $\quad \max_n \|X_n\| = \sqrt{\underbrace{1^2 + 1^2 + \cdots + 1^2}_{m}} = \sqrt{1+m}$

and $(\min_n y_n W_f^T X_n)^2$ is "closest" data to f which mean W_f and X_n have minimum inner product, since that $Z_+(X) = Z_-(X)$, where -0.5 be W_0 (threshold) of W_f $\therefore$ $W_f^T X_n = -0.5$ where $Z_+(X_n) = Z_-(X_m)$

As we know $W_{t+1} \leftarrow W_t + y_t X_t$, where y_t is sign and X_t is binary numbers 0 or 1, for every component W_i of W_f :

~~if~~ $W_i \in \mathbb{Z}$ $\forall i = 1, 2, \cdots, n$ and $W_0 = -0.5$

$$\therefore \|W_f\| = \sqrt{\underbrace{(-0.5)^2}_{W_0} + \underbrace{(1)^2 \cdots + (-1)^2}_{d}} = \sqrt{\frac{1}{4} + d}$$

d : 1 or -1 for minimize Euclidean distance

$$\therefore (\min_n y_n W_f^T X_n)^2 = \frac{1}{4} \quad \therefore T \leq \frac{(\max_n \|X_n\| \cdot \|W_f\|)^2}{(\min_n y_n W_f^T X_n)^2} = \frac{(\sqrt{m+1})^2 \cdot (\sqrt{\frac{1}{4}+d})^2}{\frac{1}{4}}$$

$$= (4d+1)(m+1)$$

#

6. pf.

Let $\{(x_n, y_n)\}_{n=1}^{N}$ where $x_0 = 1$ forming $x_n = (1, x_n^{orig})$, and

$\{(x_n', y_n)\}_{n=1}^{N}$ where $x_0' = -1$ forming $x_n' = (-1, x_n^{orig})$ have same particular sequence $n(t)$, $t = 1, 2, \cdots$ and both with $w_0 = 0$

Assume w_{PLA} and w_{PLA}' are solution of $\{(x_n, y_n)\}_{n=1}^{N}$ and $\{(x_n', y_n)\}_{n=1}^{N}$

$$w_1 = w_0 + y_{n(0)} \begin{bmatrix} 1 \\ x_{n(0)}^{orig} \end{bmatrix} \quad \text{keep of } x_i,\ i = 1, 2, \cdots, \qquad w_2 = w_0 + \begin{bmatrix} y_{n(0)} + y_{n(1)} \\ y_{n(0)} x_{n(0)}^{orig} + y_{n(1)} x_{n(1)}^{orig} \end{bmatrix}, \quad \cdots$$

$$\therefore w_{PLA} = w_T + y_{n(T)} \begin{bmatrix} 1 \\ x_{n(T)}^{orig} \end{bmatrix} = \begin{bmatrix} y_{n(T)} + \cdots + y_{n(0)} \\ y_{n(T)} x_{n(T)}^{orig} + \cdots + y_{n(0)} x_{n(0)}^{orig} \end{bmatrix}$$

$$\because x_n' = \begin{bmatrix} -1 \\ x_{n(0)}^{orig} \end{bmatrix} \quad \therefore w_{PLA}' = \begin{bmatrix} -(y_{n(T)} + \cdots + y_{n(0)}) \\ y_{n(T)} x_{n(T)}^{orig} + \cdots + y_{n(0)} x_{n(0)}^{orig} \end{bmatrix}$$

Since $y_{n(t)} w_t^T x_{n(t)} = y_{n(t)} w_t'^T x_{n(t)}' \quad \forall\ t = 1, 2, \cdots$

hence $\text{sign}\left(y_{n(t)} w_{PLA}^T x_{n(t)}\right) = \text{sign}\left(y_{n(t)} w_{PLA}'^T x_{n(t)}'\right)$

$\therefore$ w_{PLA} and w_{PLA}' are equivalent.

7.

pf. Known original upper bound is $\frac{\max_n \|X_n\|^2}{\left(\min_n y_n \frac{W_f^T}{\|W_f\|} X_n\right)^2} \geq T$ (before normalization)

If $W_0 = 0$, T is total number of mistake corrections

$$\|W_T\|^2 \leq \max_n \left\| \frac{X_n}{\|X_n\|} \right\|^2 * T = \max_n \|Z_n\|^2 \cdot T \qquad Z_n = \frac{X_n}{\|X_n\|}$$

we known $1 \geq \frac{W_f^T W_T}{\|W_f\| \|W_T\|}$, since and $W_f^T W_{t+1} = W_f^T (W_t + y_{n(t)} \frac{X_{n(t)}}{\|X_{n(t)}\|})$

$\therefore W_f^T W_{t+1} \geq W_f^T W_t + \min_n y_n W_f^T Z_n > W_f^T W_t$

since $W_f^T W_T \geq T \cdot \|W_f\| \rho_z$, $1 \geq \frac{W_f^T W_T}{\|W_f\| \|W_T\|} \geq T \cdot \frac{1}{\sqrt{T}} \rho_z = \sqrt{T} \rho_z$

$\therefore 1 \geq \sqrt{T} \rho_z \quad \therefore T \leq \frac{1}{\rho_z^2}$ new upper bound : $\frac{1}{\rho_z^2}$

8.

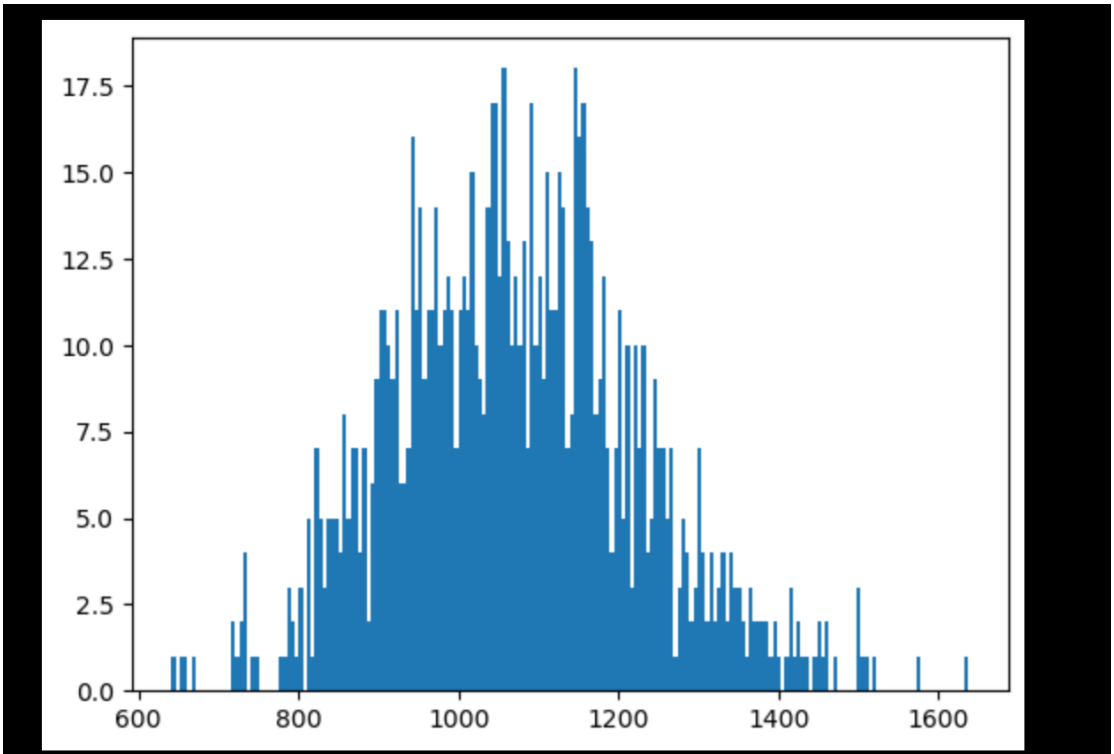
pf. Assume $D: \{(X_n, y_n)\}_{n=1}^N$ is linearly γ-separable:

$\exists W_f$ s.t. $\rho > \gamma$, that is $y_{n(t)} \frac{W_f^T}{\|W_f\|} X_{n(t)} > \gamma$ (by considering PLA is just a special case where $\gamma = 0$) $\therefore W_f^T W_{t+1} > W_f^T W_t + \gamma$

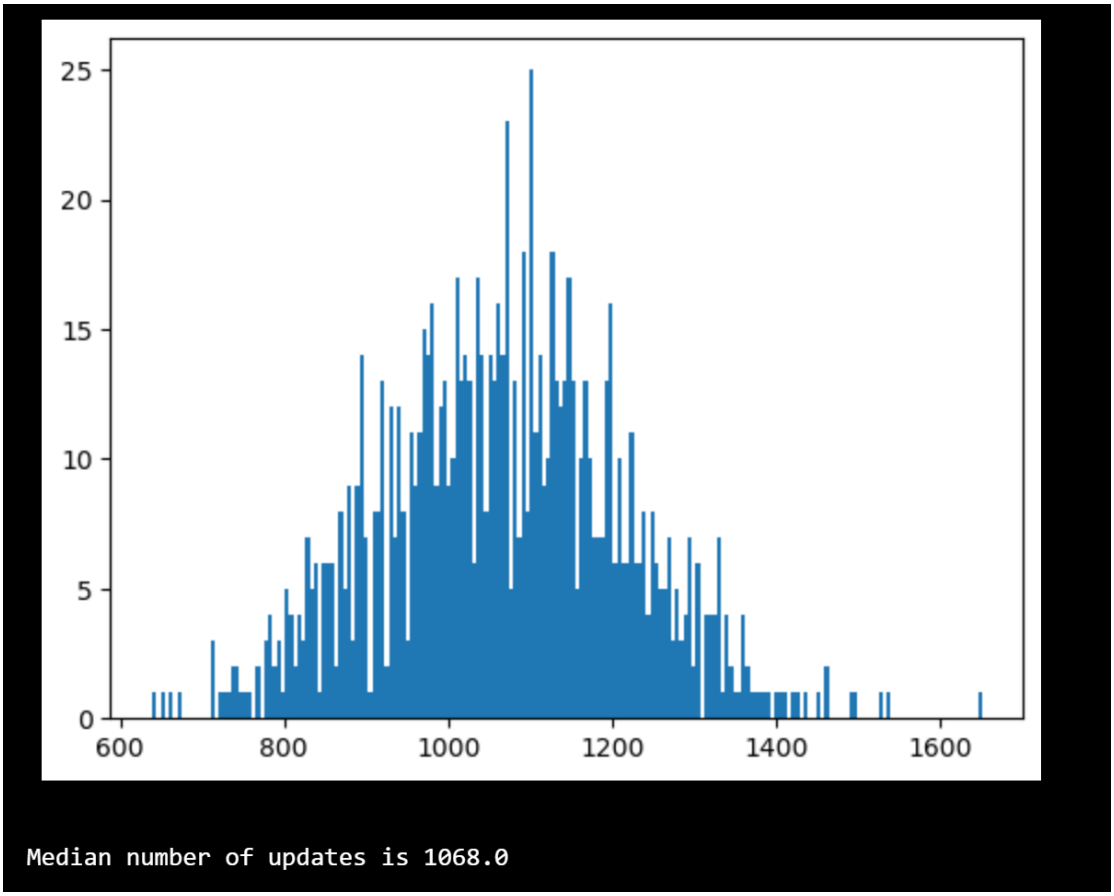
hence $\|W_{t+1}\|^2 \leq \|W_t\|^2 + 2\gamma + \|y_{n(t)} X_{n(t)}\|^2 \leq \|W_t\|^2 + 2\gamma + \max_n \|y_n X_n\|^2$

Define $R^2 = 2\gamma + \max_n \|y_n X_n\|^2$ (considering $\gamma \neq 0$) $\therefore$ PAM halts in T steps, where $T \leq \frac{R^2}{\rho^2}$ #

9

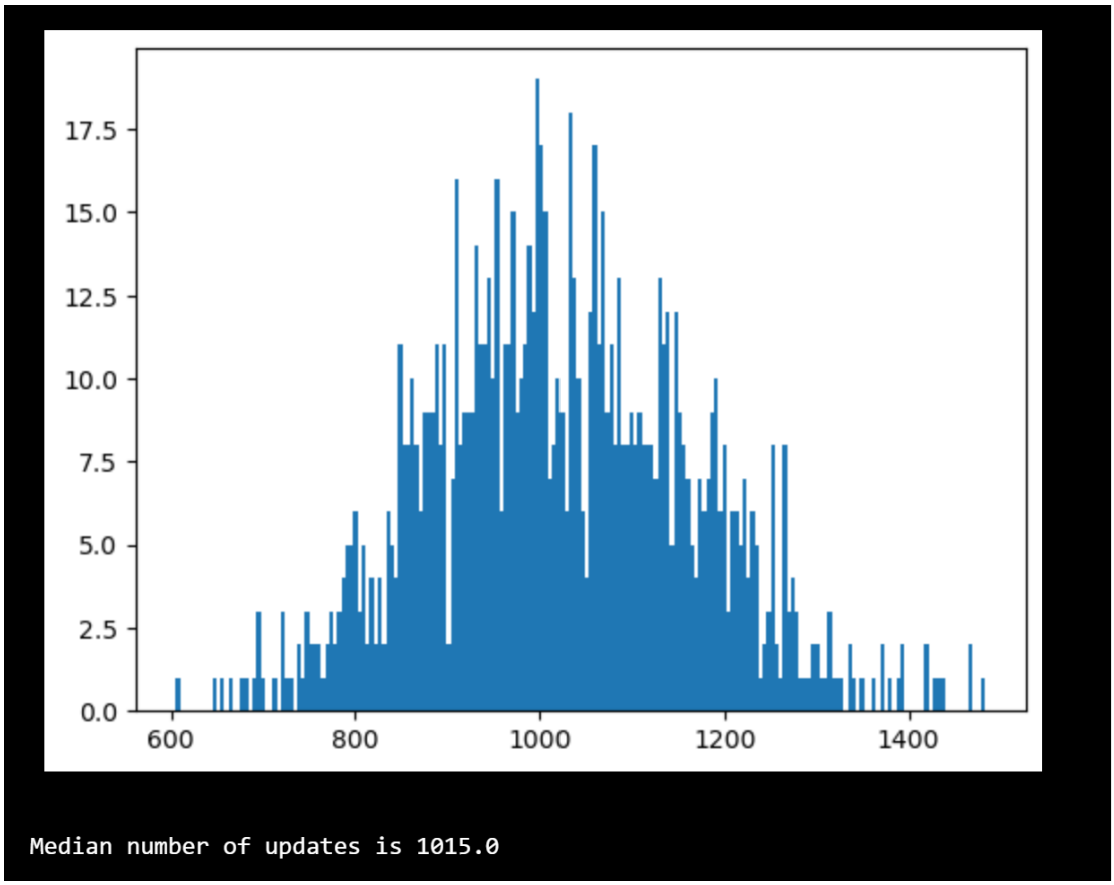

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
600
800
1000
1200
1400
1600
Median number of updates is 1066.0

10


25
20
15
10
5
0
600
800
1000
1200
1400
1600
Median number of updates is 1068.0

中位數幾乎與 9.沒有太多差異，可能是因為即使都乘 11.26，但相對尺寸仍差不多所經過的 t 差不多的關係

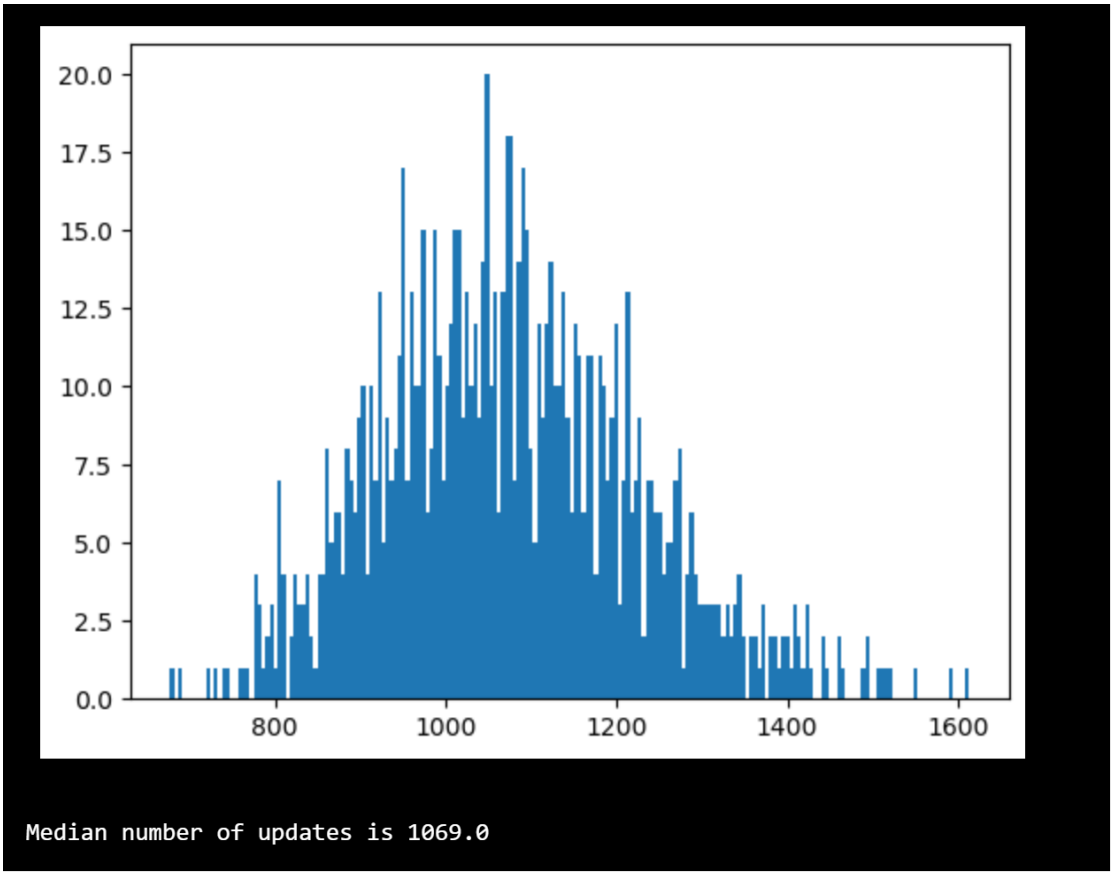
11



到這裡中位數就有很顯著的下降，這有可能是因為我們提高了 w0，無意識中影響到它逼近的步驟了。

12

與 9. 10.幾乎一樣，表示即使重複同樣的步驟，改變仍幾乎沒有。


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
800
1000
1200
1400
1600
Median number of updates is 1069.0

13

不一定 視情況而定