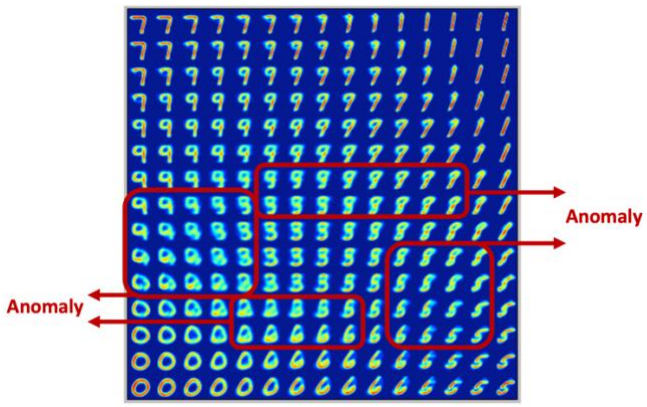
Q0: Your student ID and name. You MUST to upload your answers before 6/23 23:59.
A0: 106306062 資管三 郭宇雋

Q1: What are the pros and cons of anomaly detection and misuse detection?
A1:
Intrusion Detection 可以分為 Anomaly Detection 與 Misuse Detection 兩種。
Anomaly Detection 是透過白名單來定義好人(正常)的行為，而 Misuse Detection 則是用黑名單來定義壞人(攻擊)的行為。
Anomaly Detection 是以”異常”作為判斷是否為攻擊的主要指標，其假設為”只要異於正常就是攻擊”，因此 Anomaly Detection 的優點是可以抓出以前沒看過的攻擊。但也因為用”異常”作為判斷指標，因此缺點就是許多”異常但並非攻擊”的事件就會被錯誤判斷，False Positive 會較高。
Misuse Detection 是透過將已知的攻擊的特徵與行為寫成 ”Rule”，並禁止這些攻擊行為的發生，因此 Misuse Detection 的優點是只要以前有出現過類似的攻擊手法，則 Misuse Detection System 就可以偵測的出來。然而相對的，缺點就是若是某攻擊為新的攻擊手法，則 Misuse Detection System 就會偵測不出來。

Q2: Describe how to use an autoencoder to detect an anomaly? Probably you need draw a graph.
A2:
因為 Autoencoder 的過程是先透過 Encoder 將 input 做降維，投射到一個能夠更精簡描述資料的 latent space，而後再透過 Decoder 做還原。因此若是今天有一個 input，他跟先前我們用於訓練 Autoencoder 的 inputs 是類似的，則該 input 照理說會被還原得很好。反之，若是今天有一個 input，他跟先前用於訓練 Autoencoder 的 inputs 差距很大，則該 input 照理說就會沒辦法被還原得很好。因此我們可以建立一個用正常資料來訓練的 Autoencoder，而後當有一個新的 input 進來時，如果 Autoencoder output 是一個相似的圖片，則我們就可以判斷該 input 為正常，但若是 output 與原本的 input 不相似，則我們就會認定該 input 為異常。因此，根據 Autoencoder 還原後結果的好壞，我們就可以做到 Anomaly Detection。

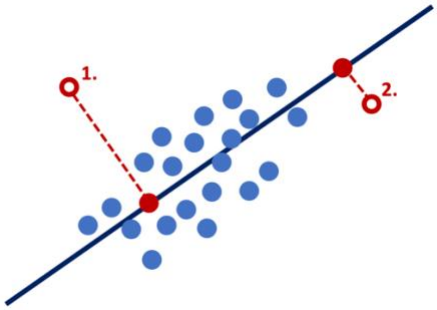


Q3: Describe how to use PCA to detect an anomaly? Probably you need draw a graph.

A3:

因為 PCA 是將原始資料轉換至新的空間 ($Z = X*W$)，當然也能從轉換後的空間再轉換回原始空間 ($\bar{X} = Z*W^T$)。因此，我們可以透過先將原始資料 X 做 PCA 轉換，而後再將其轉換回原始空間，得到 $\bar{X}$，並計算 X 與 $\bar{X}$ 的距離，距離越大表示越有可能為 anomaly。

下圖的空心紅點為欲做預測的 data point，而實心紅點表示該點經過 PCA 轉換並 reconstruct 回原始空間的結果，虛線即為距離。若以下圖來看，則點 1 較點 2 更有可能是 anomaly data。



Q4: Using one-hot encoding to encode an article (having lots of words) is impractical. What else method can you adopt?

A4:

可以使用 Bag of Words 編碼，此方法會計算各字詞在文章中出現的次數。然而，這個方法會忽略字詞出現的先後順序關係。

因此，可以改使用 Word Embedding 的方式來對文章進行編碼。Word Embedding 的原理是建立詞向量(Word Vector)，並定義向量中的每個維度各自對應到的字元為何，再將句子中的字元轉換為向量，最後將其結合成矩陣。如此的做法可以保留原先字詞出現的順序，且矩陣運算也較為方便。

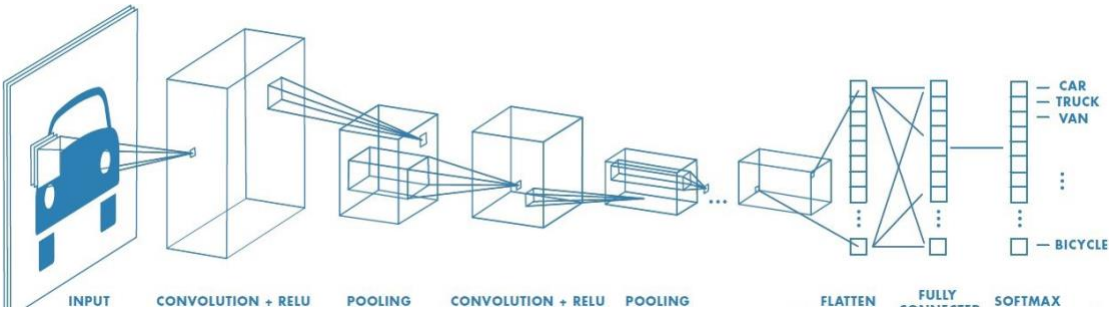
然而，若是字詞過多，則向量的長度會太長，且因為該矩陣為稀疏矩陣，因此可能會浪費許多儲存空間。

有鑑於此，可以改使用 Word2vec 的方式做編碼，這個方式的原理是透過 Continuous Bag of Word (CBOW)與 skip-gram 的方式來產生詞向量。其中 CBOW 是透過類似填空題的方式去訓練給定上下文，能夠預測輸入的字詞的模型，再用該模型編碼得到含有語意的詞向量，同時維度也能降低。

而在實務上，若是欲編碼的文章內容與 wiki pre-trained 的詞向量(w)差距不大，則可以直接使用別人(google) pre-trained 好的詞向量來用即可。若是差距過大才需要自己去抓文章作詞向量。

Q5: For an CNN (shown below, note that the second last layer is ”FULLY

CONNECTED"), please describe (1) what is the purpose of first part layers (from input to FLATTEN), and (2) what is the purpose of the last few layers (FLATTEN to SOFTMAX).



A5:

(1)

Convolution + Relu 的目的是提取圖片中的某種"特徵"，而該層的 kernel 數量就是該層會從圖片中提取(尋找)的特徵數量，而 Kernel size 表示的是該特徵的大小，而 Relu 為 Activation Function 的一種，其目的為將原先的"線性關係"轉換到"非線性"，也可以解釋成做"空間扭曲"。此外，Pooling 的做法是依序在在圖片的某範圍內的幾個格子中，挑選最大(最小)的格子做為代表，其餘格子刪除，因此可以做到類似"降維"的概念，甚至還能夠做到部分的去噪。

因此整體來說，first part layers 的過程就是透過 Convolution 不斷的在圖片中尋找不同的特徵，並透過 Pooling 做降維與去噪，以減少計算時間。而這幾層的目的就是找到 input 的圖片中，最能夠做到精準分類的特徵(最有代表性的那些特徵)。

(2)

Flatten Layer 的目的是將原先四個維度的矩陣資料攤平，以方便餵進接下來的 Fully Connected Layer。而 Fully Connected Layer 的目的則是找出某些能夠很好的進行圖片分類辨識的"觀點"。最後的 Softmax Function 為 Activation Function 的一種，用於做多元分類的輸出，因為 Softmax 會將輸出結果轉換為類似"機率"的概念，因此該層的所有 neuron 可以解釋成該 input 屬於各類別的機率，且這些機率的總和為 1。最後只要取這些機率中的最大值，即表示對該圖片的預測類別。

因此整體來說，last few layers 的過程就是將多維矩陣攤平，並餵入 Fully Connected Layer 去尋找適合的"觀點"，最後透過 Softmax 做多元分類的輸出結果。

Q6: Tell my why the above "FLATTEN" layer can be viewed as a representation of the original INPUT.

A6:

因為在 FLATTEN Layer 得到的結果是原本的 INPUT 透過許多 Convolution 不斷訓

練而尋找到的”特徵”結果，因此可以保留住原本 INPUT 中可用於分類的”重要特徵”。也因此，FLATTEN Layer 可以被視為是原本 INPUT 的特徵呈現。

Q7: Why do we need Min-Max Scaling or Z-score Normalization? How can them help the training process?

A7:

Normalization 可以幫助加快訓練速度 (加快收斂)。因為在資料集中，不同的變數(欄位)若是 range 不相同，則在做 Gradient Descent 時，不同方向下降的速度會不一致，導致需要花費較多時間才能走到終點。因此若是有先做 Min-Max Scaling 或是 Z-score Normalization，則可以幫助加快訓練過程。

此外，對於 Linear Regression 等 ML 應用來說，若是資料中有單位不相同的變數，或是 range 差距很大的變數(欄位)，例如一個是 0~1，另一個是 0~1000，則在預測上，較大的那個變數會對結果預測產生較大的影響，但這是不正確的，因為我們並無法確定較大的那個變數較為重要。因此先做 Min-Max Scaling 或是 Z-score Normalization 就可以避免這個問題。

Q8: How do you know your model overfit? How do you prevent from overfitting in a neural network?

A8:

可以先將原始資料切分成 training data 與 validation data，而後觀察 training data 與 validation data 在訓練過程中，accuracy 的變化(可以畫成折線圖觀察)，基本上 training data 的 accuracy 會不斷提升，然而，若是發現 validation data 的 accuracy 開始呈現下降的趨勢，則表示 model 發生 overfit 的問題了。而若是只能從 model 結果來觀察，則可以觀察是否發生 training data 的 accuracy 很高，但 validation data 與 testing data 的 accuracy 卻不高的情形，如果有該現象，則也表示 model 很有可能發生 overfitting。

在 Neural Network 中，可以透過設定 Early Stopping 的方式來防止 overfitting，也就是透過設定說如果對於 validation data，在特定數量的 epochs 內 accuracy (或是 loss)都沒有明顯改善，則停止訓練。

此外，在 Neural Network 中，還可以透過添加 Dropout Layer 的方式來防止 overfitting，其原理類似 ensemble learning 的概念，也就是每次 epoch 的訓練中，都隨機將其中幾個 node 丟掉(設為 0)，如此可以避免 model 因為只透過少數幾個特徵(觀點)去做預測，而發生 overfitting。

此外，在 Neural Network 中，overfitting 的發生也有可能是因為 Layer 的數量過多，或是 neuron 的數量過多，導致訓練出來的 model 過於複雜，因而發生 overfitting。因此，有時候也可以考慮減少 Layer 或是 neuron 的數量。

最後，還有其他方式也可以防止 overfitting，包括對 Image Data 使用 Data Augmentation 來增加資料量，常見有旋轉、縮放等等；或是使用 Regularization

在 Loss Function 添加 Penalty 的方式來避免因為某些 weight 過大導致發生 overfitting，常見有 L1 與 L2 Regularization。

Q9: What is the difference between noise and anomaly? How to remove them?

A9:

Noise 大部分時候是 unwanted 的，也就是對資料分析沒有幫助；然而 Anomaly 常常是資料分析主要關注的點，例如詐欺偵測、入侵偵測等等。

而 Noise 可以透過降維 (如 PCA、Autoencoder) 來去除。此外，對於圖片資料，也可以使用 Denoising Autoencoder (DAE) 來訓練能夠去除 Noise 的 Autoencoder。

而對於 Anomaly Data，也可以透過如第二題與第三題(Q2,Q3)所提及的 PCA 與 Autoencoder 這兩個方法，來判斷是否為 Anomaly 並做後續處理(ex.移除等等)。

此外，亦可透過 KNN 等方式，去對距離設定一個 threshold，若超過該 threshold 則認定為 Anomaly，再對該資料做後續的移除等處理。

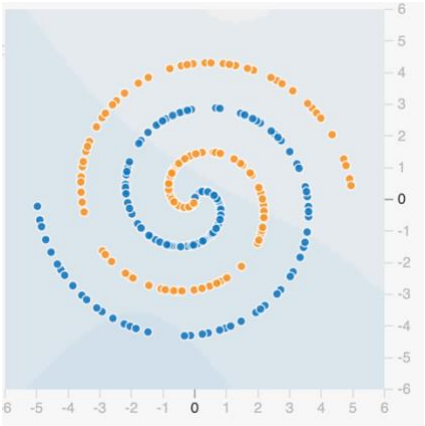
Q10: What is the kernel of a CNN?

A10:

CNN 中的 kernel 表示的是某個"觀點"，或是也可以說是某個"特徵"，因此 Kernel 的主要目的就是從圖片中萃取出某些特徵。Kernel 基本上是一個帶有某個數值(帶有某特徵)的小矩陣，每次訓練時該 Kernel 會依序在圖片上移動，並在對應的小區塊上面做每一格的對應相乘運算，並將該運算的結果加總。因此可以知道，若是圖片上的某個區塊與 Kernel 矩陣的數值很接近，則其運算結果就會很大；而若是某區塊與 Kernel 矩陣的數值差距很大，則其運算結果就會很小，因此可以從這個結果得知圖片中是否存在與該 Kernel 很相似的特徵。

此外，對於某層 Convolution Layer，kernel 數量就是該層會從圖片中提取(尋找)的特徵數量，而 Kernel size 表示的是該特徵的大小。

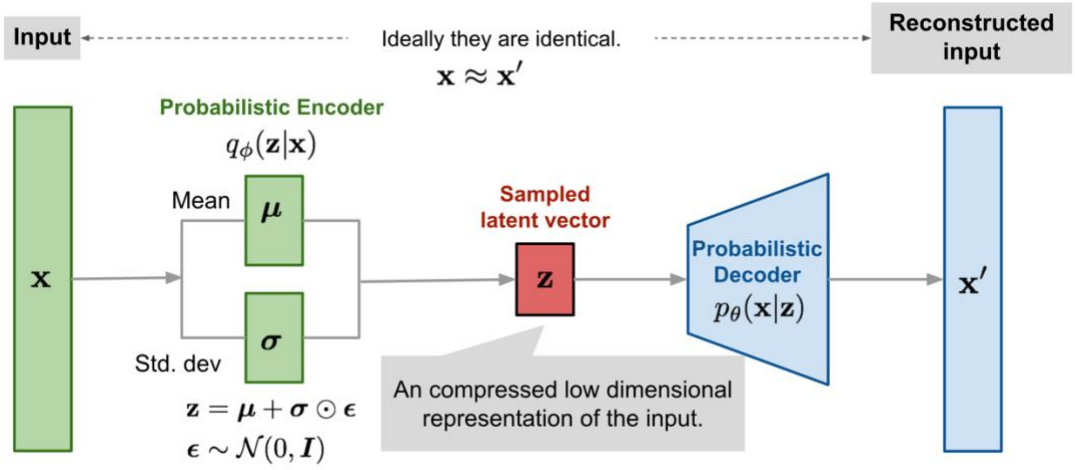
Q11: Why a neural network can classify orange points and blue points? Give us some high-level explaination.



A11:

因為 Neural Network 在做的事情其實就是平移、旋轉以及空間扭曲，其中平移與旋轉是透過每一層的 w (觀點)來達成，而扭曲則是透過 Activation Function 來達成，好讓我們可以將線性不可分的問題轉換為線性可分(Linear Separable)，並將資料點的類別分開。因此對於題目的圖片，Neural Network 可以透過多次的平移、旋轉與空間扭曲，來將這些看似線性不可分的 orange points 與 blue points 的空間不斷扭曲，最後將其分開。

Q12: What is the epsilon in VAE? Why it is a normal distribution N(0, I)?



A12:

VAE 的最終目的是為了能夠產生新的 X，因此首先要確保 latent space 要均勻分布，且不能 overfit，如此才能產生好的 output。而為了要達到這個目的，就得透過 Epsilon。

Epsilon 是遵循 normal distribution 分佈的資料點，加入 epsilon 類似加入 noise，目的是透過 normal distribution 的抽樣使分布均勻，進而得到較好的 output。

而 Epsilon 為 normal distribution 的原因是因爲想透過加入 Epsilon，使 latent space 的產出(z)形成均勻的分佈，亦即 latent space 內的所有點都可以被選到，因此為了達到這個目的，要讓加入的 Epsilon 為 Normal Distribution 才可以。