Machine Learning Final Project

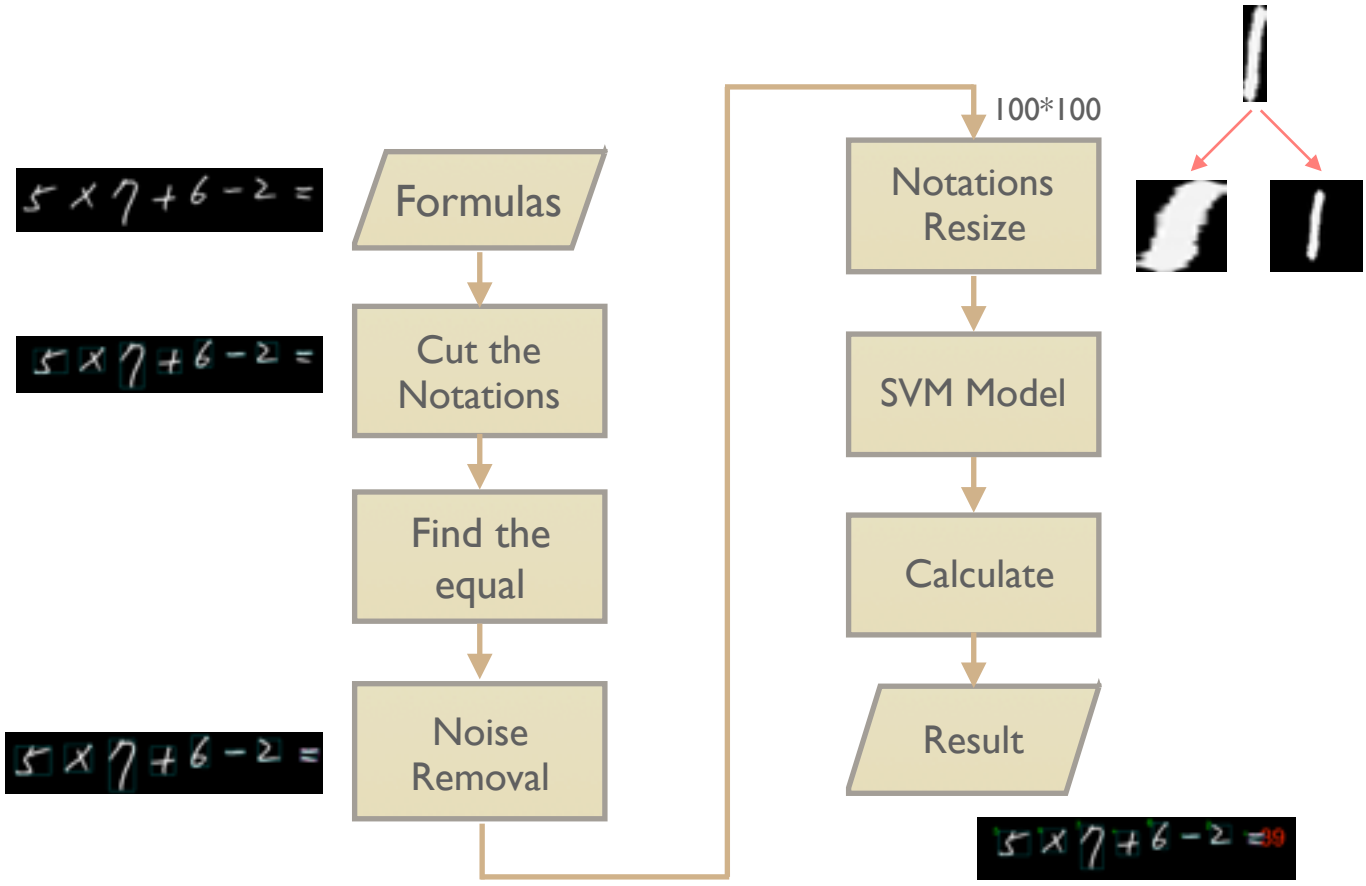
Handwritten Formula Competition

A041540 蔡博丞
B041525 吳儀萱

Introduction

- Training Data : White_digit ➤ Bold & Normal
- Test Data : White_digit ➤ Normal
- Classifier : SVM Model (Radial basis function : C-SVM)

Flow Chart



Notation Extract

IDEA I

- Segmentation

為了將文字、運算符號一個一個切割出來, 我們將整張formulas上的資訊先對其橫座標軸取Histogram, 將Histogram值連續不為0的部份切割出來, 即可得到每一個字元的切割.(水平及垂直兩個方向都要做, 才能確切地切出適當的範圍)

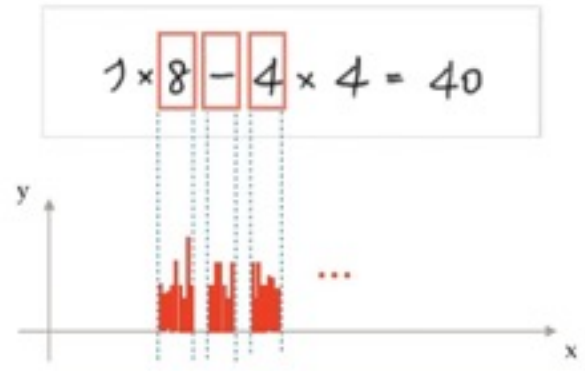


Fig.1-1 Histogram diagram

但是，這個方法當遇到下面的情況就會失敗，因為括號和數字寫得太近時，就會把他們兩個一起切割出來，這個問題還在研究中。

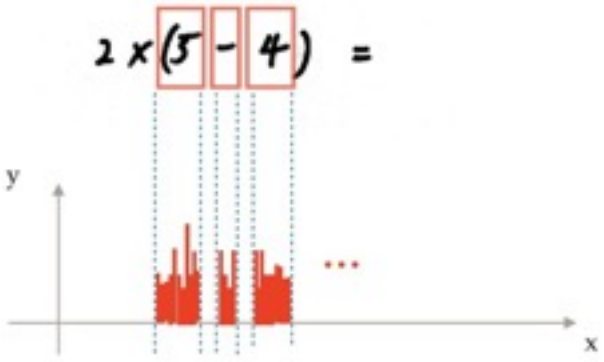


Fig.1-2 Histogram diagram

IDEA II

- Cut the Notations
  - Method : Connected Components
    - ➢ Connected Components :
      先找出圖中的所有points, 圈選出points有相連的部份,
      則為我們所需要的Notations.

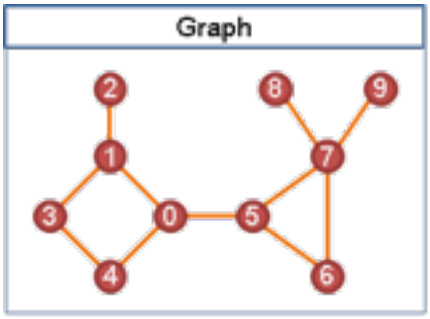


Fig.2-1
Connected Components diagram

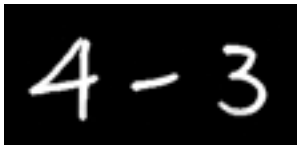


Fig.2-2 Formula

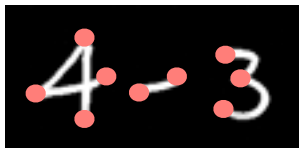


Fig.2-3
Formula's Points

- Noise Removal
  - Method : 刪除小於block size的部份
  - Block Size : 11*9

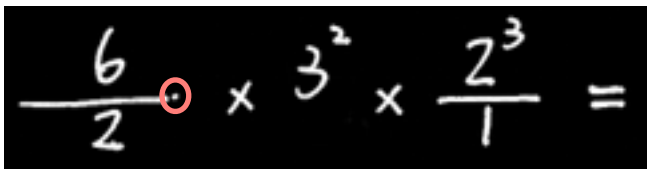


Fig.3-1 有Noise的Formula

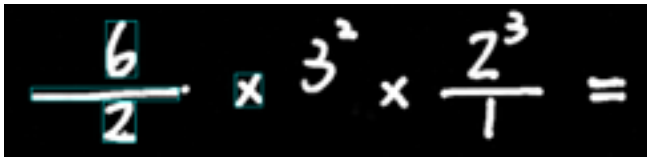


Fig.3-2
Notation extract after noise removal

✦ Find the Equal

- STEP 1：挑選被圈選出類似－(減號)的所有block (e.g.分號 / 等號)

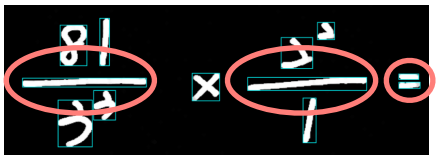


Fig.4-1 Formula

- STEP 2：計算每個被挑選出的候選block的長寬比 (Ratio=長/寬),
  將Ratio>1.5 且 length<60 的候選block留下.

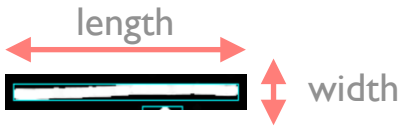


Fig.4-2
A semicolon cut from the block

- STEP 3：計算留下的候選block的前後座標點. (fig.3-3 A,B,C,D).

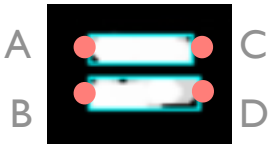


Fig.4-3
Equal coordinate diagram

- STEP 4：將兩兩候選block作前座標點及後座標點相減,
  (fig.3-3 AB & ~~CD~~) 把相減值<15的留下.
  ❊ 若只做前座標點相減, 可能會出現錯誤的情形. (fig. 4-4)
  而後座標點相減也加入考慮則可避免此問題. (fig. 4-5)

- STEP 5：將留下的候選block結合在一起, 即為"=".

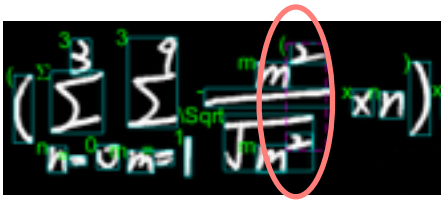


Fig.4-4

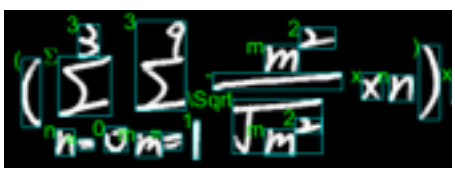


Fig.4-5

- Formula Analysis

  先將Formula分成三種類別, 分別是 ”Hard”、“Normal”、“Easy”,
  接著利用下方流程圖的步驟將”Hard Formulas”進行簡化.

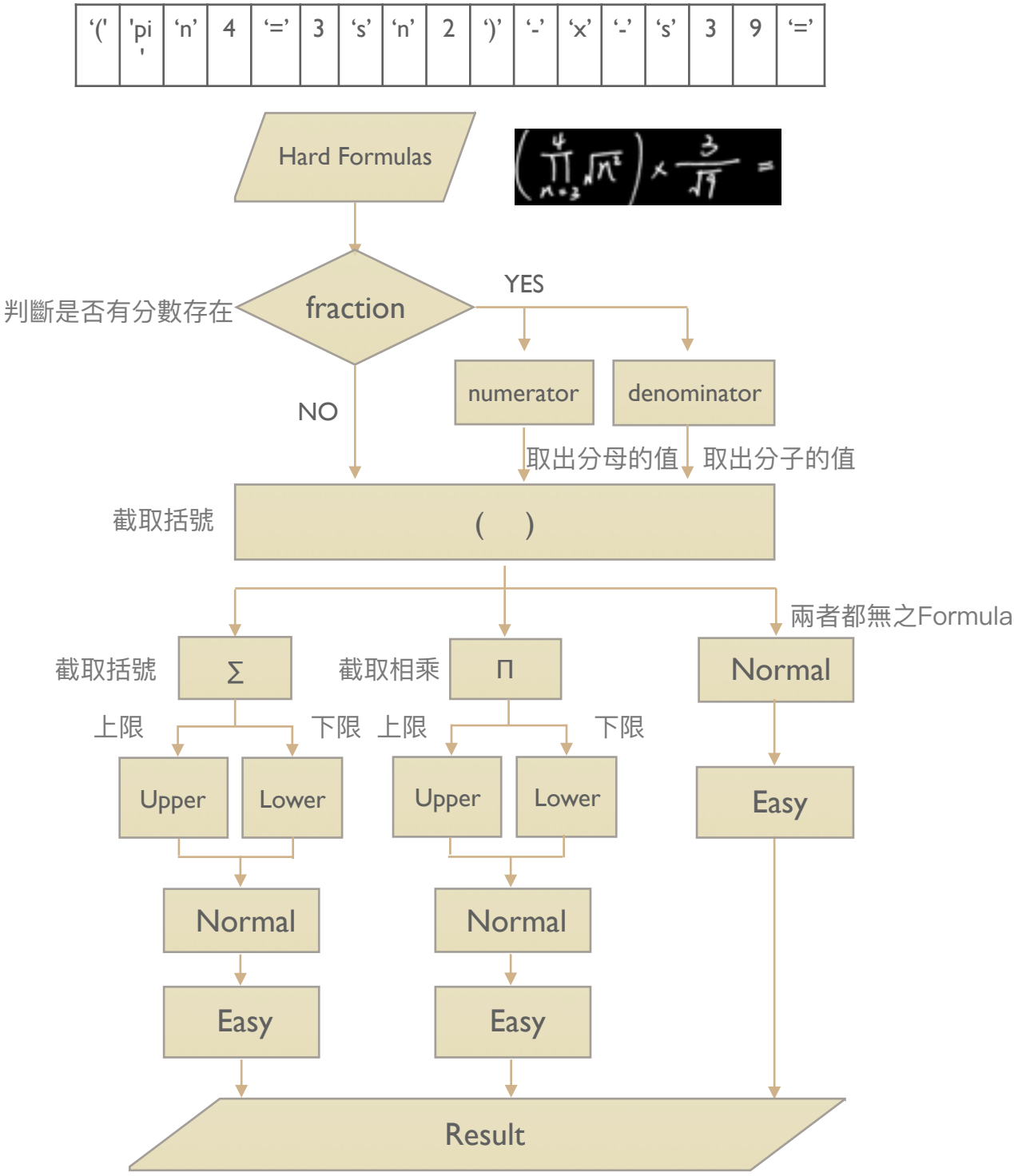


Fig.4-6 Formula analysis diagram

而簡化成”Normal Formula”後, 則需要使用下面流程圖中的步驟繼續簡化至”Easy Formula”.

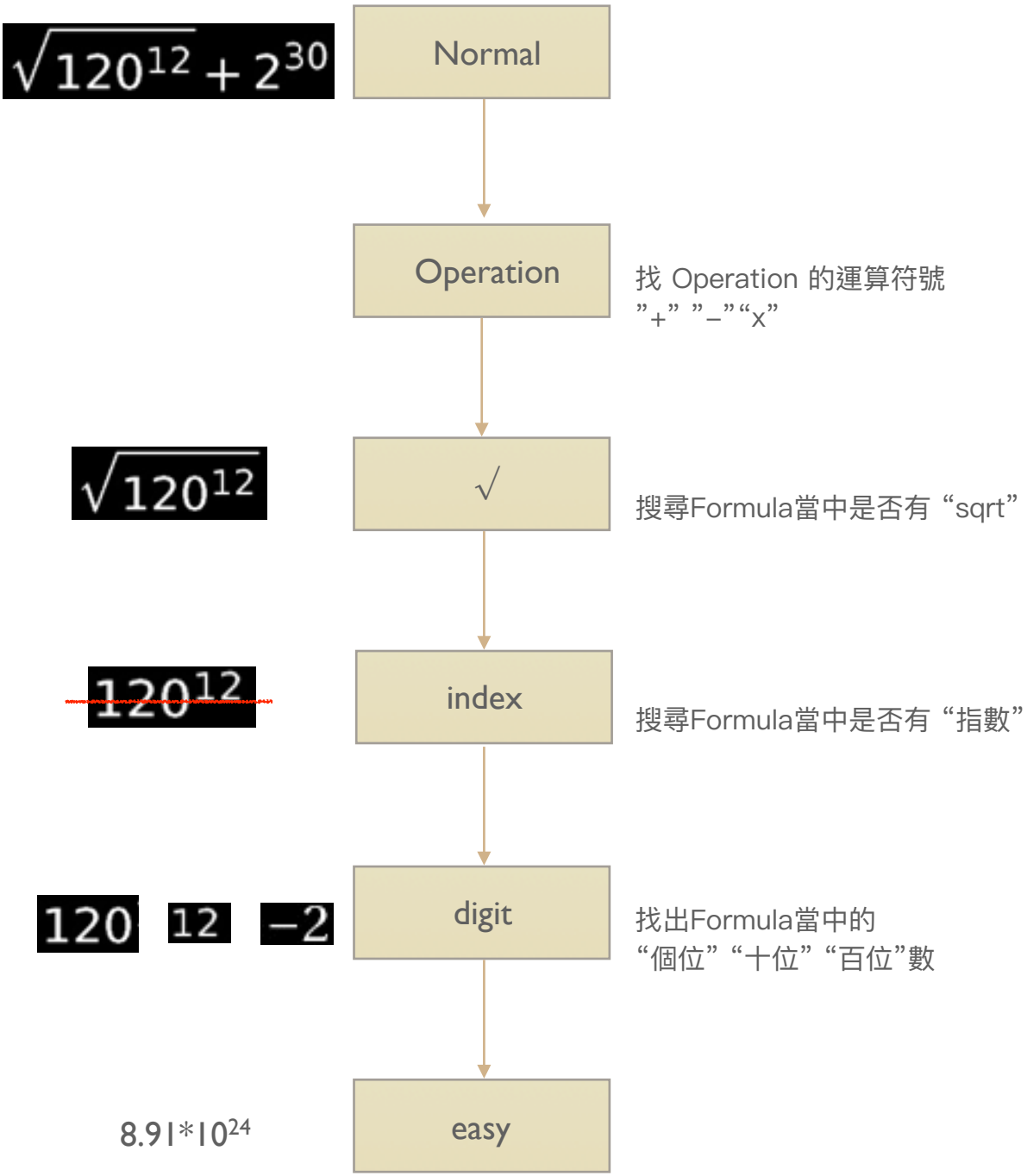


Fig.4-7 Formula analysis diagram ||

Training Database

- 增加 Database：擷取每個label當中的些許data,
  – 1, 2, 3, 4,…,9, (,), −, =, π, Σ : rotate +15 degree & −15 degree
  – +, × : rotate +5 degree & −5 degree
  （因為 + 旋轉過多會變成 × , × 亦是, 所以旋轉角度較小）

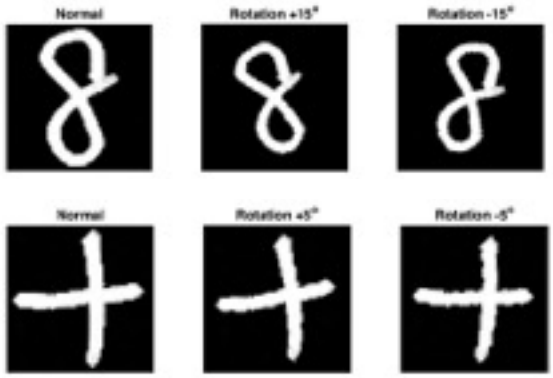


Fig.5-1
未旋轉及旋轉後的Notations

- Training data 總量：
  960(number) * 21(label) + 600(fraction) = 20760
  ❊ 由於fraction容易辨識錯誤, 所以有增加一些其他角度旋轉後的 data來增強fraction的訓練.

- Dimensionality Reduction: Principal Components Analysis (PCA)
  由於Data量較大, 以至於維度很高, 會使訓練過於繁複及時間過長,
  因此使用PCA的演算法來進行降維的動作.

- Validation Data： 6700 筆
  從提供的Notations中, 每個label都截取一部份當作Validation data,
  進而得以計算Model的準確率.

SVM Model

- Classifier : Radial basis function ➢ C-SVM
- Model I 的Training data有多增加fraction項及旋轉項,
  因此Model I 比Model II 分類更準確.

	Training Data	Validation Data	Parameter	Dimension	Accuracy (Validation)
Model I	20760	6700	-s 0 -t 2 -g 0.002 -c 250 -e 0.4	29	95.38%
Model II	20160	6400	-s 0 -t 2 -g 0.005 -c 250 -e 0.4	15	93.89%

* e: epsilon

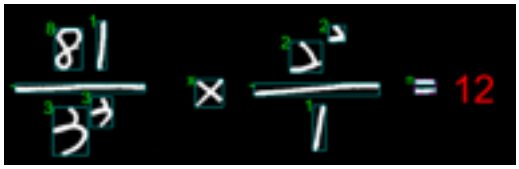


Fig.6-1 (a) Model I

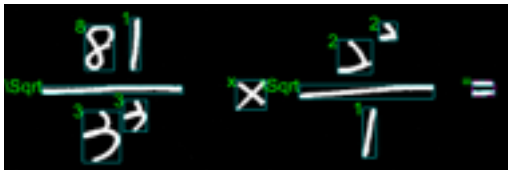


Fig.6-1 (b) Model II

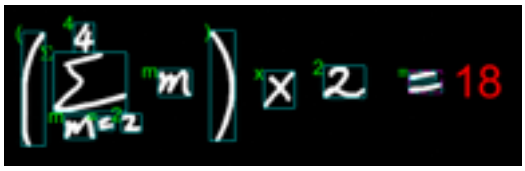


Fig.6-2 (a) Model I

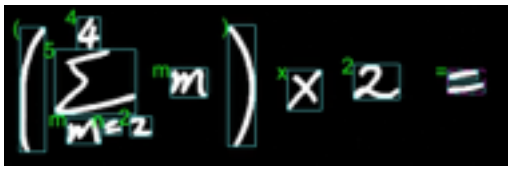


Fig.6-2 (b) Model II

- 比較：從Fig.6-1可以看出(a)可以辨識出fraction, 則(b)無法.
  Fig.6-2中(a)可以辨識出sigma, 則(b)無法.

- 分析：由於 Model I 使用較多的Training Data來做訓練,
  因此分類的效果也比較準確.

Result

	LEVEL 1 ACCURACY	LEVEL 2 ACCURACY	LEVEL 3 ACCURACY	LEVEL 4 ACCURACY	LEVEL 5 ACCURACY
Model 1	33%	40%	30%	13%	15%

Future Work

Fault		
1	(	1
2	2	)
3	3	)
4	+	4

- 缺失：有幾個label容易互相辨識錯誤, 以及字跡當中同一符號但筆畫有空隙的部分容易無法辨識正確 (有些連圈選出Notation都有些問題).
- 改進：原本使用的方法(connected component)會無法正確的切出數字出來，主要是因為他是用pixel和pixel之間的相對位置所找出來的，然而這會造成同個數字，但寫得比較不連續一些，就會造成他們分離，而變成兩個不同的代表數字。
- 因此我想嘗試運用region的資訊去掃整張圖，像是Faster R-CNN就是一種我想嘗試的方法，利用它object detection的model來確切的框出數字出來。