

Image Processing Final Report

107703005 資科四林庭安 107703040 資科四王豐惟

June 17, 2022

1 Introduction

1.1 參加比賽

肺腺癌病理切片影像之腫瘤氣道擴散偵測競賽：運用物體偵測作法於找尋STAS

1.2 比賽介紹

「腫瘤氣道擴散(spread through air spaces, STAS)」是近年新發現的肺腺癌病理特徵，指的是腫瘤細胞從腫瘤的邊緣沿著肺泡空腔，往附近的正常肺臟組織擴散。STAS是肺癌不良預後的獨立影響因子，對手術後的復發風險有重大影響，已是目前肺腺癌病理檢查的重要項目。由於辨識玻片中的STAS需花費大量人力及時間，此比賽希望透過人工智慧方法自動找尋病理切片中的STAS，除了節省醫師的時間外也縮短判讀時間，為患者爭取寶貴的治療時機。

本比賽提供STAS訓練資料集1053張，公開資料集(Public images)131張，以及私有資料集(Private Images)184張。我們需利用訓練資料集進行模型的訓練，接著預測公開資料集和私有資料集中的STAS，將結果上傳至比賽網頁可獲得公開資料集的預測結果，而私有資料集的預測結果則需等比賽截止日後才可得知。

1.3 物件偵測

本比賽希望透過物件偵測方法來進行STAS的偵測。物件偵測為影像處理中的重要一環，該方法透過邊界框(bounding box)框住模型預測出的目標物，而每個邊界框會預測框住物體的類別、位置以及該物體為目標物的機率。物件偵測方法可用於進行多類別的偵測，而本比賽偵測的物體為單一類別，故結果只會是STAS；邊界框的位置預測會以座標的形式表示，不同方法所表示的座標形式不同，我們所使用的方法是以邊界框中間點x座標、邊界框中心點y座標、邊界框長度以及邊界框寬度來標示目標物，且邊界框中心點x座標和邊界框寬度會除以圖片寬度，而邊界框中心點y座標和邊界框長度會除以圖片長度，以正規化的形式表示。

判斷物件偵測預測結果好壞的評估標準為AP (Average Precision)，其公式如下：

$$AP = \sum_t^{T-1} \left(P_t + \frac{(P_{t+1} - P_t)}{2} \right) (R_{t+1} - R_t) \quad (1)$$

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{n} \quad (3)$$

$$TP = \sum_i^n 1\left\{ \max_{j \in \{x \in N | 1 \leq x \leq \hat{n}\}} IOU(b_i, \hat{b}_j) \geq IOU_{th} \right\} \quad (4)$$

$$FP = \hat{n} - TP \quad (5)$$

其中， P_t 及 R_t 為在第t置信度(confidence)域值的精確率(precision)和召回率(recall)，T為精確率和召回率集的大小， b 為基準真相(ground truth)框，而 \hat{b} 則為預測框。

由上式可知，計算AP的方法為先透過真陽性(Truth Positive, TP)和偽陽性(False Positive, FP)算得本次預測結果的精確率和召回率，接著便可得到PR曲線(Precision-Recall Curve)，由式(1)可知，AP可透過計算本次預測結果的PR曲線底下的面積所求得，且AP越大，代表預測結果越好。物件偵測計算TP的方法可參考式(4)，其透過計算邊界框和與其相鄰最近的基準真相之間的IoU，並與我們所設定的IoU域值比較，若其值大於域值則將此物體歸類為TP，若小於域值則為FP。本比賽所使用的評估標準為AP^{.50}，意即IoU域值為0.5。

2 Method

本比賽我們將透過使用Yolov5[1]方法來進行物件的偵測，此方法為one-stage的物件偵測方法，因此他具有訓練速度與偵測速度快的優點，在比賽時可以節省時間，進行更快速地驗證。我們所使用的訓練環境為Colab所提供的Tesla T4 GPU，記憶體大小約為15GB，訓練圖片大小為900，batch size為35。訓練資料集中訓練資料為737張，驗證資料為211張，測試資料為105張。

Figure 1為本次比賽的訓練資料集、公開資料集和私有資料集，可以觀察到資料集中不同資料在顏色上的差異性大，且STAS的形狀變化亦為不規則，因此為了使模型預測的結果更穩健，我們將透過影像前處理(image preprocessing)和資料擴增(data augmentation)方法來針對影像的顏色和目標物形狀進行處理。連結為我們實作的程式碼：<https://github.com/aoi-mitoru/ImageProcessingFinal>

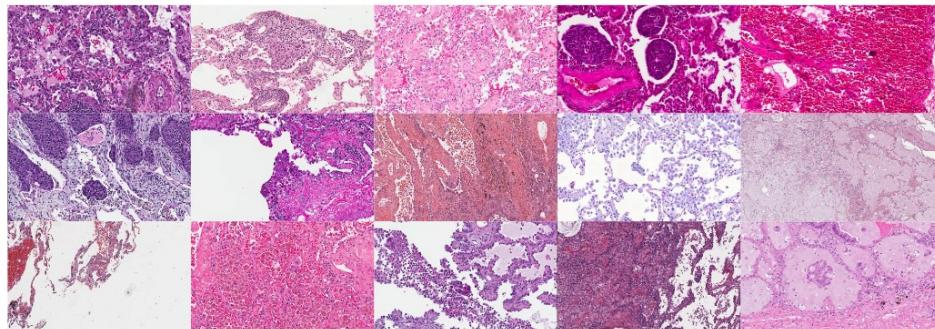


Figure 1: 資料集

2.1 Image Preprocessing

1. 影像預分割

由於Yolov5在訓練圖片時會將圖片壓縮成正方形，我們認為比賽提供的資料集圖片的長寬比例差異很大(942x1716)，在壓縮後可能影響影像品質，使訓練效果不如預期。因此希望透過將原始圖片切割成六份，使圖片長寬比較為接近(471x572)。分割方法如Figure2所示。

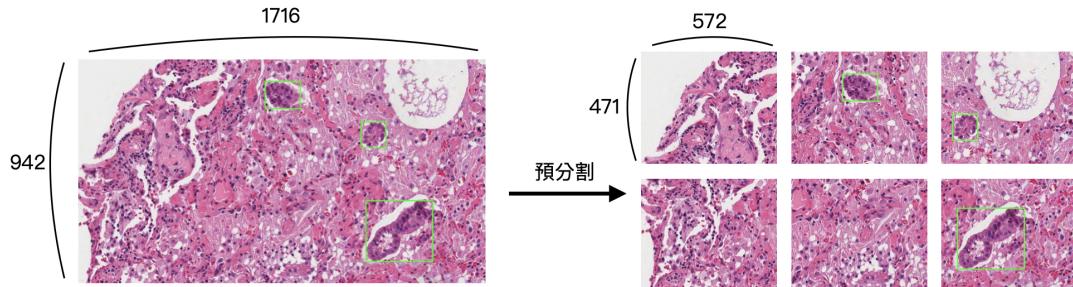


Figure 2: 預分割

2. 影像標準化

由於我們觀察到比賽方提供給我們的資料集間（訓練資料集、公開資料集以及私有資料集）資料的差異性大，導致最後驗證結果不如預期，因此我們希望能降低資料集間的差異性，使得預測結果能更準確。老師在期末報告發表時建議我們進行標準化(Standardization)的前處理方法，因此我們嘗試透過對一張圖片的H、E、D三個通道分別進行Z分數標準化(Z-Score Standardization)，使其平均值為0，標準差為1，讓圖片與圖片間的HED三個通道的分布一致（關於HED色彩空間可參考Section2.2.1）。接著由於進行標準化後圖片內的像素值過於相近，因此我們分別對H和E通道乘上0.7，而D通道則乘上0.2，我們認為在此類型的染色上D通道相較於H和E通道的重要性較低，因此將D通道乘上較少比例。標準化前後對比圖如Figure3所示。

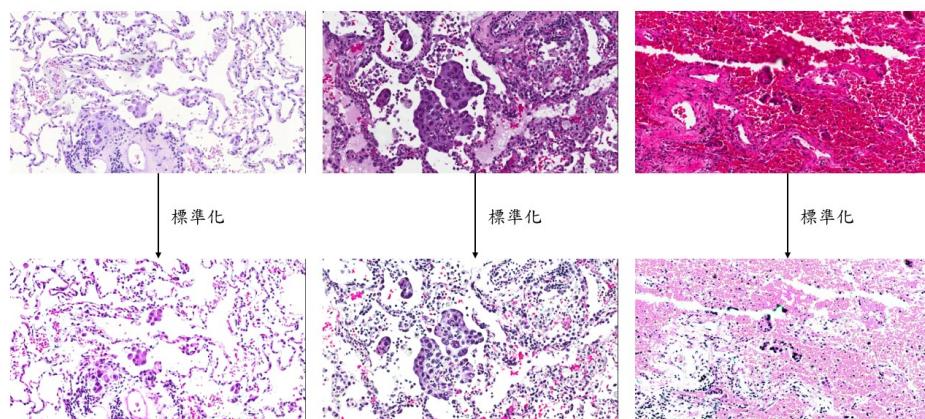


Figure 3: 影像標準化

2.2 Data Augmentation

本比賽使用的資料集以蘇木精-伊紅染色 (hematoxylin and eosin stain, H&E stain) 方法為玻片進行染色。由於玻片影像來自各個不同的實驗室，導致玻片的染色情況不盡相同，再加上腫瘤細胞有各種形狀，因此我們希望透過資料擴增增加資料集的多元性，使模型能訓練得更穩健。

腫瘤影像資料具有Stain Invariance、Morphology Invariance與Artifact Invariance三種特性，可以此為基礎進行資料擴增[3]。

1. Stain Invariance

我們透過修改蘇木精與伊紅顏色通道的強度來模擬各種不同的H&E染色影像。首先將影像從RGB色彩空間轉換至HED色彩空間，接著使用從連續型均勻分布中取出的隨機因子 α_i 與偏差項 β_i 分別修改H、E、D三個通道，最後再將影像從HED色彩空間轉換回RGB色彩空間以製造不同的H&E染色影像。

比爾-朗伯定律 (Beer-Lambert law) 是光吸收的基本定律。其物理意義為，當一束平行單色光垂直通過某一均勻非散射的吸光物質時，其吸光度與吸光物質的濃度成正比。

$$OD_i = -\log \frac{I_i}{I_{0,i}} = -A_{ci} \quad (6)$$

其中 OD_i 為吸光度、 I_i 為透射光的強度、 $I_{0,i}$ 為入射光的強度、 A 為吸光物質的濃度、 c_i 為吸光係數，下標 i 用以表示RGB三個顏色通道之一。

基於此定律，我們可用OD向量描述單一染劑在R、G、B三個顏色通道的OD值。蘇木精染劑的OD向量為 $[0.18, 0.20, 0.08]$ 、伊紅染劑的OD向量為 $[0.01, 0.13, 0.01]$ 、DAB染劑的OD向量為 $[0.10, 0.21, 0.29]$ [2]。將三種染劑的OD向量組合在一起可組成一代表OD轉換值與HED染色空間間線性關係的矩陣，如下所示。

R	G	B	
0.18	0.20	0.08	Hematoxylin
0.01	0.13	0.01	Eosin
0.10	0.21	0.29	DAB

(7)

為了使每個染劑在吸光度的貢獻相同，我們將每個OD向量除以其長度以形成標準化後的矩陣。

R	G	B	
0.65	0.70	0.29	Hematoxylin
0.07	0.99	0.11	Eosin
0.27	0.57	0.78	DAB

(8)

因此一個以OD值表示的像素 y 可以以下式表示：

$$y = xM \quad (9)$$

x 為一個代表單一像素染色量的1乘3的向量， M 為標準化後的OD矩陣。為了求得 x ，我們將 y 乘以 M 的反矩陣，即為下式所示：

$$x = yM^{-1} \quad (10)$$

我們將一張 $P \in R^{M \times M \times 3}$ 的RGB玻片影像reshape成 $P \in R^{N \times 3}$ 後，可利用標準化後的OD矩陣 M 轉換成HED色彩空間

$$S = -\log(P + \epsilon)M^{-1} \quad (11)$$

其中 $S \in R^{N \times 3}$ 為轉換為HED色彩空間的玻片影像， ϵ 為避免 \log 內為零的正偏差項。

接著透過隨機的線性轉換改變各染劑的強度

$$S'_i = \alpha_i S_i + \beta_i \quad (12)$$

$S'_i \in R^{N \times 3}$ 為改變染劑強度後的影像、 α_i 取自均勻分佈 $U(1-\sigma, 1+\sigma)$ 、 β_i 取自均勻分佈 $U(-\sigma, \sigma)$ ，論文中建議的 σ 值為0.05。

最後再次利用OD矩陣將影像從HED色彩空間轉換回RGB色彩空間即可得到隨機增強各染劑強度的玻片影像。

$$P' = \exp(-S'M) - \epsilon \quad (13)$$

由於我們發現 σ 值為0.05時得到的影像會出現偏向藍色、綠色等Public資料集與Private資料集中不會出現的顏色，因此我們嘗試將 σ 值調整為0.02以便顏色更符合資料集。如Figure4與Figure5所示

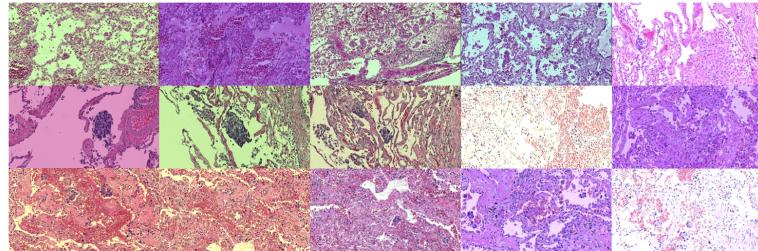


Figure 4: $\sigma=0.05$ 之隨機染劑增強資料擴增

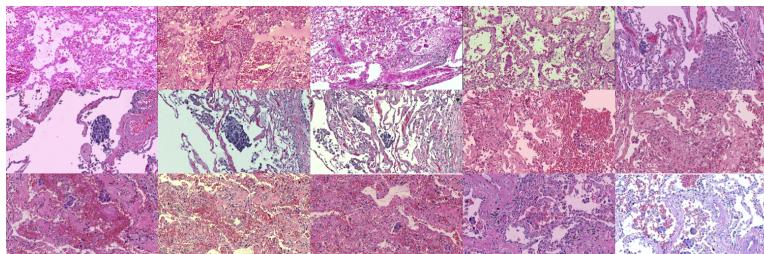


Figure 5: $\sigma=0.02$ 之隨機染劑增強資料擴增

2. Morphology Invariance

由於腫瘤細胞具有不同形狀，我們利用旋轉與翻轉擴增資料。如Figure6所示。

3. Artifact Invariance

我們利用高斯濾波器（Gaussian Filter）模擬玻片影像失焦的情況，並隨機從0.5到1.5的範圍內取出一值作為模糊半徑(radius)。如Figure6所示。

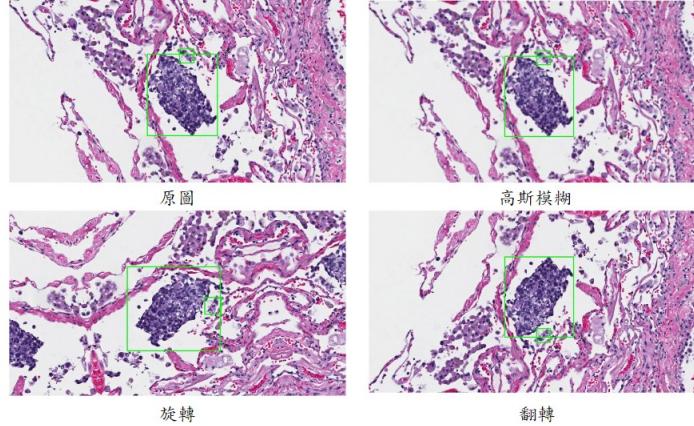


Figure 6: 旋轉、翻轉與高斯模糊

3 Results

preprocessing	data augmentation	Test AP ^{.50}	Public AP ^{.50}	Private AP ^{.50}
x 預分割	x	0.794	0.726	x
標準化	x	0.751	0.722	0.636
x	stain ($\sigma = 0.05$)	0.750	x	x
x	stain ($\sigma = 0.02$)	0.779	0.780	0.661
x	stain ($\sigma = 0.05$) +artifact	0.771	0.782	0.657
x	stain ($\sigma = 0.05$) +morphology	0.787	0.744	0.655
		0.794	0.743	0.679

Table 1: 各項實驗結果

從Table1可看出，在Public資料集中成績最好的方法為 $\sigma=0.02$ 時之隨機染劑增強資料擴增，而Private資料集中成績最好的方法為 $\sigma=0.05$ 時之隨機染劑增強資料擴增加上旋轉與翻轉資料擴增。由於Public資料集中的影像間差異較小，而Private資料集中的影像在顏色、亮度與腫瘤形狀上都有很大的變化。因此由Public資料集的結果無法準確判斷Private資料集的結果。

值得一提的是，原以為預分割後成績應比預分割前更好，不過結果卻不如預期。我們推測原因為影像進行分割後，位在分割處位置的腫瘤會因為被切割而遺失資料，導致腫瘤的訓練資料相較分割前缺少許多，如Figure7所示。

由於使用標準化對影像進行前處理是在比賽結束後才嘗試的方法，我們只能以自行切割訓練資料集所得到的Test資料評估此方法。由成績來看此方法的表現並不好，不過我們無法在Public資料集

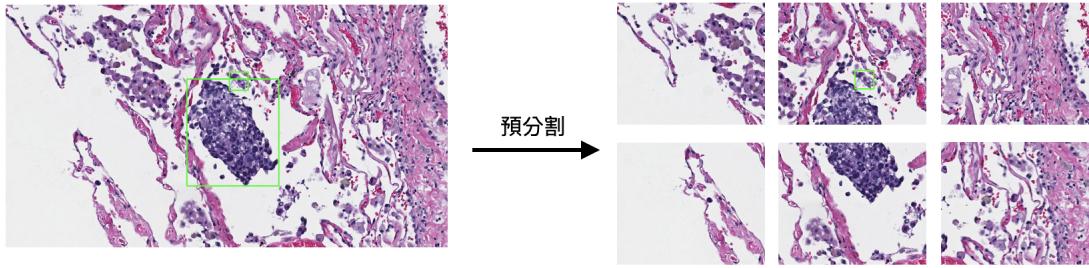


Figure 7: 預分割

與Private資料集上確認其結果。

4 Conclusion

在本比賽中我們採用Yolov5方法進行物件偵測，並嘗試透過前處理和資料擴增方法提升準確率，但或許是因為我們對於模型訓練的方法上鑽研不足，結果不如預期的好，因此未來若有機會再參與類似的比賽，除針對資料本身進行處理外，我們將比較各類型模型，並採用合適的訓練方法。

References

- [1] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu 于力, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, October 2020.
- [2] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.
- [3] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.