# 國立高雄科技大學

# 資訊工程系碩士班

# 碩士論文

## 在有限資料中進行影像分類

## **Image classification with limited data**

研究生：冷宛圻

指導教授：鐘文鈺 博士

中華民國 **110** 年 **8** 月

# 在有限資料中進行影像分類

# Image classification with limited data

研究生：冷宛圻
指導教授：鐘文鈺 博士

國立高雄科技大學
資訊工程系碩士班
碩士論文

A Thesis
Submitted to
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Computer Science and Information Engineering

Aug 2021
Kaohsiung, Taiwan, Republic of China

**中華民國 110 年 8 月**

# 國立高雄科技大學（建工校區）研究所學位論文考試審定書

資訊工程系 碩士班

研究生 涂宛圻 所提之論文

論文名稱(中文): 在有限資料中進行影像分類

論文名稱(英/日/德文): Image classification with limited data

經本委員會評審，符合 碩士 學位論文標準。

學位考試委員會

召 集 人 王子元 簽章

委 員 王子元 陳沁瑾

鍾子敏

指導教授 鍾子敏 簽章

系所主管 陳沁瑾 簽章

中華民國 110 年 8 月 11 日

保存期限：永久

**Image classification with limited data**

**by**

**Wan-Chi Tu**

A Thesis
Submitted to
Institute of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Computer Science and Information Engineering

Aug 2021
Kaohsiung, Taiwan, Republic of China

Approved by:

王子元 (Tzi-Yuan Wang)

Wanju Chang

Ju-chin chen

Advisor: : Wanju Chang

Department Chairman : 國立高雄科技大學電機與資訊學院資訊工程系主任 陳洲瑾

# 在有限資料中進行影像分類

學生：涂宛圻　　指導教授：鐘文鈺 博士

國立高雄科技大學資訊工程系碩士班

## 摘要

機器學習在許多領域上的表現已經非常接近人類的行為，但大多都需要大量的數據才能有效地學習，獲得好的效果。而對於某些領域來說，要取得大量數據是非常困難，或者成本太高，因此有少樣本學習的演算法被提出，希望能解決這項障礙。雖然這些模型在圖像單純的 Omniglot 數據集上有出色的表現，但在圖像較複雜的 miniImageNet 數據集上表現卻不佳。所以我們提出一整合方法，先利用邊緣偵測取得圖像中的前景，再執行物件偵測，框選影像中的主體，並擷取出來，以此讓主體特徵更明確。之後透過修改的關聯網路（Relation Network）進行影像辨識學習，原本是以串接方式將特徵圖連接起來計算數據間的相似度，這裡我們修改成相減來計算。訓練後，可以藉由 sample set 與 query set 的關係得分來對影像進行分類。

關鍵字：少樣本學習、關聯網路、物件偵測、邊緣偵測、選擇性搜索

# Image classification with limited data

Student：Wan-Chi Tu Advisor：Dr. Wen-Yu Chung

Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology

## ABSTRACT

The performance of machine learning in many fields is close to that of humans, but most training models require a lot of data to learn effectively. In some fields, it is difficult to obtain a large amount of data, for example, the cost of collecting data is high, so few-shot learning models are proposed to overcome the obstacle. Although these models perform well on the Omniglot dataset with simple images, they perform poorly on the miniImageNet dataset with more complex natural images. Therefore, our method is to use Edge-detection to obtain the foreground in the image, and then use Object-detection to frame the subject in the image, which by extracting the subject to make its characteristics clearer. Finally, we use a modified relation network as our training model by replacing the concatenation process by subtraction. After training, the image can be classified by the relation score between the sample set and the query set.

Key words: Few-shot learning、Relation Network、Object Detection、Edge Detection、Selective Search

# 目 錄

# 表目錄

# 圖目錄

# 一、緒論

## 1.1 研究動機

影像辨識在開發時，如果遇到了訓練樣本數不足的問題，會發生辨識效果不佳的情況。一般在做機器學習時，是需要有大量訓練數據才能有效學習，尤其在影像辨識的任務上依賴更多的數據。但在某些領域，例如野外的物種辨識或調查，難以取得大量數據。所以要想藉由機器學習作為任務的輔助，資料量是首先就會碰上的困難點。

新國民食魚教育，是由國立高雄科技大學漁業生產與管理系劉仁銘教授，所執行的一項宣導漁業永續資源重要性的計畫，目的是希望消費者在市場挑選可食用魚時，能購買已達到成熟期的魚。因為達到成熟期的魚最少有產過一次後代，可避免該魚種被大量捕食後，沒有留下足夠數量的後代繼續繁衍的疑慮。因為智慧型手機的普及，該計畫希望利用可以在手機上運作的影像辨識 APP，作為推廣計畫的媒介，當消費者操作 APP 拍攝時，APP 會辨識畫面中所拍攝到的魚種，接著根據辨識結果，獲得該品種成熟期應達到的標準魚長。

所以 APP 同時會執行物件偵測，將畫面中偵測到的魚及參考物，例如：硬幣，做比例計算，以此獲得該魚的實際體長。若計算得到的長度不小於該魚種的標準長，則代表這條魚已達到成熟期，至少有產過一次後代；反之則代表尚未成熟。最後消費者會在手機畫面上看到這條魚的品種名稱，以及是否建議購買的反饋訊息。

由中央研究院生物多樣性研究中心王子元研究副技師，所執行的溪哥的生態調查的計畫，也是希望能藉由影像辨識 APP，來做為調查人員在野外即可使用的辨識設備，以便於觀察本土種溪哥和外來種溪哥在同一生態域的發展，及各個種類的數量調查。

魚類照片最直接的取得來源是下水拍攝，因此必須準備可在水中拍攝的設備。除此之外，第一個計畫例子也可到市場找魚攤販拍攝，但魚都是堆疊在一起的，畫面會過於雜亂，且較難辨識個體長度。而第二個計劃例子可以請釣客協助提供

野生溪哥的照片，但無法要求照片背景是簡潔的。由以上兩個計畫例子可知要取得大量且背景單純容易辨識的照片是不容易的事情。

## 1.2 研究方法概述

在被提出來適合做少樣本學習的方法有遷移學習[1]、度量學習[2]、元學習[3]等，其中最常用的方法是元學習。這些方法之間是可以混用的，例如，元學習使用距離度量來查詢相似度。它們在單純的圖片中有不錯的效果[2,4,5,6,7]，但在較複雜的自然影像上表現卻不好[4,5,6,7]。所以我們的研究方法是將物件偵測技術加入基於度量學習的元學習模型，來預測圖片的類別。因為要收集到大量的樣本數很不容易，就算透過資料增強方法生成更多樣本，數量仍可能不夠，因此我們希望模型能直接學習影像主體，不考慮不相關的背景圖案，所以我們利用影像去背及物件定位，將照片上主要物件的部分保留下來，可以讓物件的特徵更明確；我們同時也修改了關聯網路，預期能提高少樣本模型預測的準確度。

# 二、相關資訊與文獻探討

## 2.1 少樣本學習

機器學習要表現出如同人類的水平，通常需要大量的數據做為學習依據，例如 ImageNet 是影像領域中常用的一個自然影像數據集，該數據集中包含兩萬多類，照片總數量高達一千四百多萬張。但實際上人類學習一個新類別時，並不需要依靠那麼多的照片進行學習，而且我們也很難收集到那麼多元的類別，或是每一類的樣本數量只可能收集到幾百張至幾千張照片。在 2015 年，Brenden M. Lake 等人提出機器學習可以像人類一樣，僅從已知一個或幾個例子中去學習的新概念[8]。在這個概念提出後，陸續有研究團隊發表的深度學習模型[2,4,9,10]表明深度學習確實能從少數幾個例子中學習。這對於不易取得大量數據的領域來說，少樣本學習[11,12]的提出，正是他們所期待的。

數據和學習是機器學習的兩個重點，因此可以將少樣本學習的方法大致分為數據方法和學習方法兩類[12]。數據方法是針對數據做處理，可能嘗試找到更明確的特徵；或是利用現有的少量數據去生成更多數據。而學習方法是改進模型的學習，方法有三種：第一種，遷移學習（Transfer Learning）[1]，利用在相關的大型數據集預訓練網路後，將知識轉移到自己的模型上；第二種，度量學習（Metric Learning）[2]，又稱為相似度學習，利用數據間的距離來計算數據間的相似度；第三種，元學習（Meta Learning）[3]，該方法與一般機器學習不同，元學習是通過大量任務訓練，然後測試學習新任務的能力。元學習是在少樣本學習問題中最常用的方法。

## 2.2 元學習

元學習又稱為Learning to Learn，意即讓模型學會如何學習。與一般機器學習以單個任務的訓練方式不同，元學習是透過多個不同任務的方式來訓練。訓練元學習需要兩個網路，一個為學習大量新任務的learner，一個為訓練learner的meta-learner。元學習方法通常分為三類，第一類是Recurrent Models [9,13]，用循環網路做為learner，例如Long Short-Term Memory（LSTM）[14]。第二類是Metric Learning[2,4,5,6,7]，使用距離函數計算樣本間的相似度。這兩類的meta-learner都

是採用一般的optimizer。而第三類：Learning Optimizers[15]，learner為一般的卷積神經網路，但利用循環網路來做為meta-learner。

我們論文是採用 Metric Learning 方法，它是利用距離函數計算兩個數據在空間上的距離，再依距離的遠近來做為兩個數據是否相似的判斷。由於 Metric Learning 方法的輸出結果只有相似或不相似，所以這方法通常是屬於二元分類的問題。

2015 年，Gregory Koch 等人提出卷積孿生網路模型[2]，他們是採用孿生網路（Siamese Network）結構的卷積神經網路來提取特徵，之後利用 L1 距離計算兩個特徵間的差距。孿生網路的結構是由兩個相同結構且共享權值的子網路組成，如圖 1。

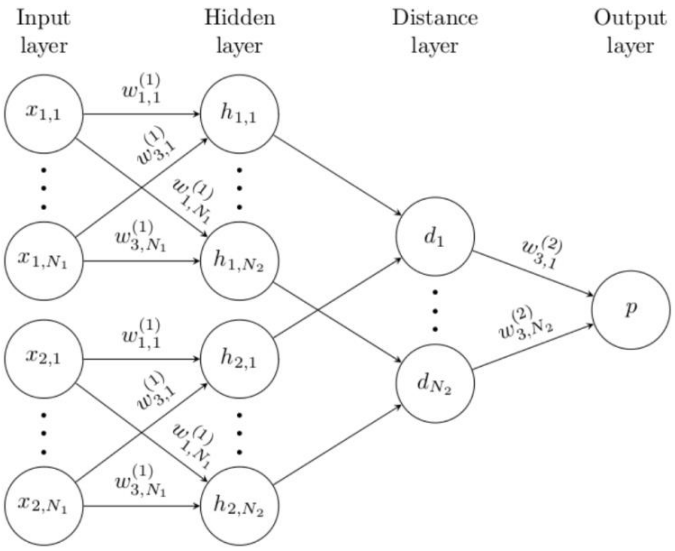

圖 1、孿生網路結構[2]。由兩個相同結構且共享權值的子網路組成的網路。

2016 年，Oriol Vinyals 等人提出匹配網路模型（Matching Network）[4]，除了利用距離函數之外，他們還借鑒了注意力機制和外部記憶的概念來建立模型，而會採用外部記憶是認為樣本應該還會跟前後的樣本有關，所以利用雙向 LSTM 來做為傳遞訊息的網路，之後再利用結合注意力機制（Attention Mechanism）的 Cosine 相似度來計算兩個特徵間的差距。

2017 年，Jake Snell 等人提出原型網路（Prototype Network）[5]，他們用卷積神經網路提取特徵，之後用平方歐式距離來計算兩個特徵間的差距。在距離函數上，他們融合了聚類概念，會先對每一類都提取樣本的平均值，每一類計算出來的平均值就會做為該類的 prototype，如圖 2。之後依照要查找的數據與哪一類的 prototype 的距離遠近，以判斷它是屬於哪一類。

2018 年，Flood Sung 等人提出關聯網路（Relation Network）[6] ，如圖 3，是用卷積神經網路分別建立了兩個模塊，*f* 模塊及 *g* 模塊。*f* 模塊提取圖片特徵後，以串接（concatenation）方式將兩個數據的特徵圖組合後，送進 *g* 模塊中學習。

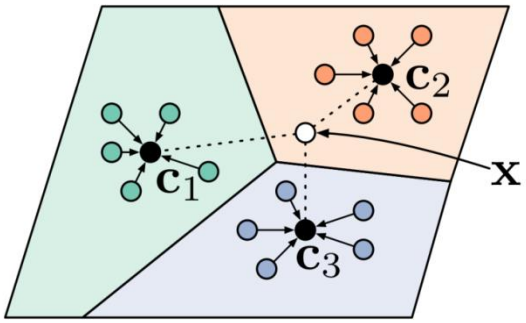


圖 2、原型網路（Prototype Network）[5]。融合聚類概念的元學習模型，圖中綠色、橘色及藍色區塊為不同類別，$C_1$、$C_2$ 及 $C_3$ 分別為該類的 prototype。X 是要查詢與哪一類相似的新樣本，它與 $C_2$ 的距離較近，所以會跟橘色區塊的樣本較相似。

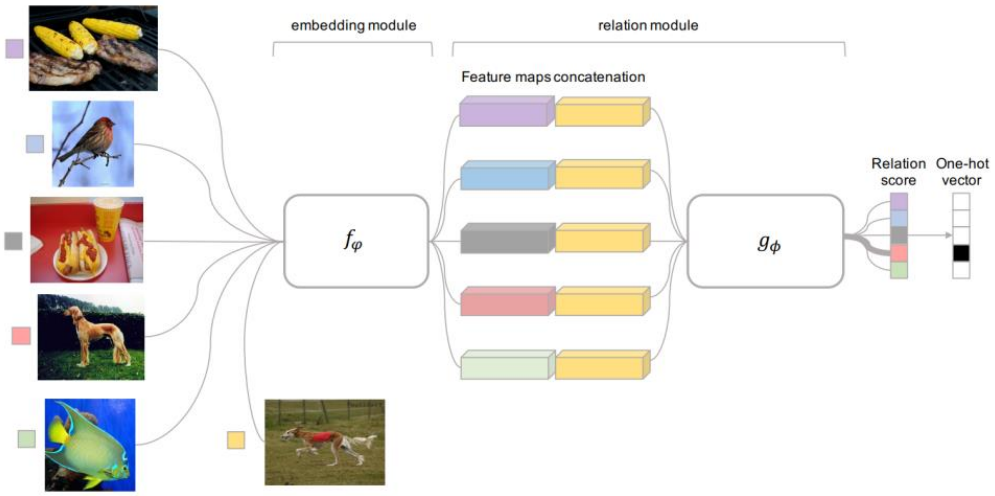


圖 3、關聯網路[6]。Embedding module 提取特徵，relation module 計算樣本間的相似度。 中間以串接方式（concatenation）將兩張特徵圖連接起來。

2018 年，Junhua Wang 等人提出基於注意力機制的孿生網路模型[7]，也是採用孿生網路結構，但是改用 Inception V4[16]來提取特徵，之後再利用結合注意力機制的歐式距離計算兩個特徵間的差距。

這些模型在圖像單純的 Omniglot 數據集上（https://github.com/brendenlake/omniglot/tree/master/python），分類的準確度都有在 90%以上[2,4,5,6,7]，但在以自然影像組成的 miniImageNet 數據集（https://github.com/vieozhu/MAML-TensorFlow-1）中，因為圖像較為複雜，所以效果並沒有那麼好，準確度只有 40%至 50%上下[4,5,6,7]。

元學習是採取學習不同任務的方式，但如果每項任務都用大量數據訓練，會非常耗費訓練時間，因此在元學習中，每項任務的訓練數據，都是以訓練集的子集合構成。N-way K-shot[4,5,6,7]是元學習方法常用來表示訓練數據集合的設定，sample set 是做為已知類別的樣本集合，query set 則是未知類別的樣本集合，也就是說要查詢 query set 裡的每個樣本與 sample set 裡的哪個類別相似。Sample set 是從訓練集中隨機挑選 N 個類別，再從被挑選出的 N 個類別中，每一類都隨機挑選 K 個樣本，此時會有 N*K 個樣本作為 sample set。而 query set 則是從這 N 個類別被挑剩下的樣本中，每一類隨機挑選部分的樣本作為 query set。Sample set 和 query set 會組合成一個任務，也就是一個 episode，如圖 4。

圖 4、5-way 1-shot 示意圖，圖片來源 miniImageNet。N=5，K=1。

## 2.3 物件偵測

物件偵測是機器學習在影像研究中的一項技術，它的目的是在影像中先尋找到感興趣的區域後，再對找到的這些區域進行分類[17]。這項技術的應用在各個產業上越來越常見，例如：行人偵測、自駕車、安全監測、醫療照護、工地安全等。我們的方法僅需要尋找感興趣的區域，並不對區域進行分類，因此以下只討論標出感興趣區域的方法。

傳統的物件偵測在尋找區域的方法是使用多尺度滑動窗口掃描整張影像，這方法可以找出所有可能有物件出現的位置，但找到的區域不一定都是有意義的，這樣就會形成冗餘的區域。但若限制它搜尋區域的數量，則可能出現找到的區域不是令人滿意的情況。

也有許多加入深度神經網路的物件偵測模型，分為區域建議法跟迴歸/分類法，如圖 5。區域建議法跟傳統的物件偵測一樣，會先尋找區域，再對區域做分類，也就是 two-stage approach。這類知名的方法是 R-CNN 系列，其中 R-CNN[18]及 Fast R-CNN[19]是利用選擇性搜索來找到感興趣的區域，Faster R-CNN[20]則提出 RPN（Region Proposal Networks）來做為候選框的提取網路。採用迴歸/分類法的則是將物件的定位及分類同時完成，也就是 one-stage approach，知名的方法有：YOLO[21]、SSD[22]。

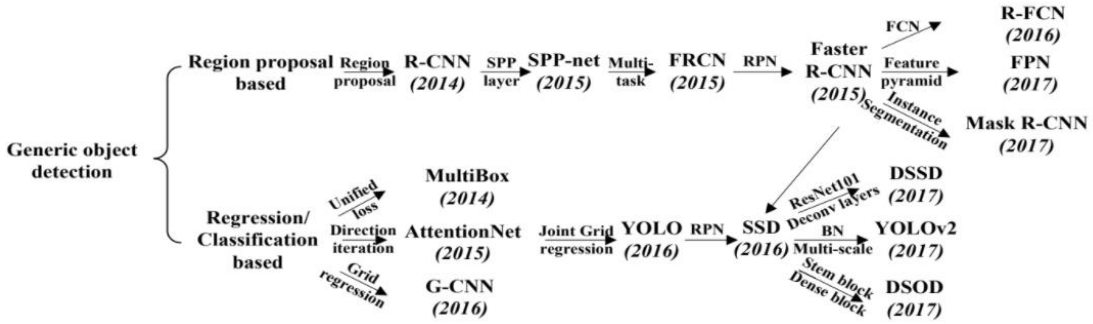

圖 5、通用物件偵測的兩種類型[17]。

### 2.3.1 選擇性搜索

選擇性搜索是階層群聚演算法（Hierarchical Grouping Algorithm）[23]，它是以圖形分割演算法[24]為基礎，然後進行階層合併來產生感興趣的區域。不過因為影像的複雜性，它無法只使用單一特徵相似度就能處理多數的情況，所以選擇性搜索是結合四種特徵相似度，分別是顏色相似度、紋理相似度、大小相似度、形狀相似度，來計算區域相似，以進行區域合併。

### 2.3.2 邊緣偵測

除了選擇性搜索，我們還利用邊緣偵測[25]來做物件定位。邊緣偵測是以像素間的變化程度來區分邊界。理想情況下，此演算法應用在影像時，可以減少數據量，過濾掉不需要的部分，保留影像中的重要結構。

除了物件定位之外，我們也利用邊緣偵測做出簡易的影像去背（Image Matting）效果。影像去背的主要目的是要將前景從影像中擷取出來，通常分為三種問題，單色去背、差異去背及自然影像去背。單色去背的背景會是單一種顏色，所以只要去除一種顏色就可以得到前景，例如：拍攝電影或是新聞報導時，背景會使用藍幕或綠幕。差異去背需要兩張影像，一張是純背景，一張是有前景跟背景，把兩張影像做相減計算就可以取得前景[26]。如果影像的背景並不是單一種顏色，也無法取得純背景的影像時，就要採用第三種的自然影像去背方法來做，這方法也是需要兩張圖，一張是 Trimap，一張是前景跟背景的原圖。Trimap 要利用原圖來產生，它是人工去劃分，將原圖上的每個像素分為前景、背景及前景背景交界處不易分辨的未知區域。有了 Trimap 後就要計算未知區域是屬於前景還是背景，以製作成遮罩圖。有了遮罩圖後，將遮罩圖跟原圖做 AND 計算，就可以取得前景。傳統作法是藉由演算法去計算未知區域的歸屬，不過用演算法處理的分離效果並沒有很穩定。2017 年採用深度學習來計算未知區域的歸屬，效果可做到分離毛髮的細緻程度[27,28]。不過因為 Trimap 不能批量的處理所有圖，它要由人工一張張的畫，會太耗費人力，所以我們用邊緣偵測取代 Trimap 來一次性的產生所有的遮罩圖。

# 三、研究方法

## 3.1 資料來源與下載

### 3.1.1 Omniglot

Omniglot 包含來自 50 種不同文化跟語言的字母表，總共有 1623 類的字母，每一類字母有 20 個不同的筆跡，圖片大小為 105*105 像素，如圖 6。可從 https://github.com/brendenlake/omniglot/tree/master/python 下載。

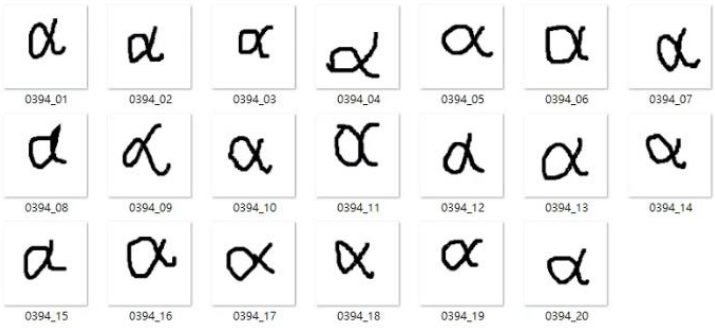


圖 6、Omniglot 數據集中字母的示意圖。圖中為希臘文字 α 的 20 個筆跡。

### 3.1.2 miniImageNet

2016 年由匹配網路的作者 Oriol Vinyals 等人[4]提出的一個小型數據集，是從大型數據集 ImageNet 中隨機挑選 100 類，每一類有 600 張不同的照片，如圖 7。本論文所使用的 miniImageNet 數據集可從 https://github.com/vieozhu/MAML-TensorFlow-1 下載[29]。

圖 7、miniImageNet 數據集中的示意圖。

### 3.1.3 新國民食魚教育計畫

台灣西南部魚市場常見食用魚類，總共 2 類，分別為花身雞魚有 7 張，如圖 8；短棘鰗有 8 張，如圖 9。這個數據集有利用拍攝不同旋轉角度的魚做為數據擴增，擴增後花身雞魚有 499 張，短棘鰗有 575 張。由國立高雄科技大學漁業生產與管理系劉仁銘老師實驗室提供。

圖 8、花身雞魚的範例照片。

圖 9、短棘鰗的範例照片。

### 3.1.4 溪哥

台灣常見的淡水魚溪哥總共有 4 種，此數據集僅收集其中 3 種，分別為台灣特有魚種的粗首馬口鱲（Opsariichthys pachycephalus），有 390 張，如圖 10；台灣常見種的長鰭馬口鱲（Opsariichthys evolans），有 328 張，如圖 11；及日本外來種的平頷鱲（Zacco platypus），有 250 張，如圖 12。由中央研究院生物多樣性研究中心王子元研究副技師提供。經過資料增強（https://keras.io/api/preprocessing/image/）後，粗首馬口鱲有 2580 張，長鰭馬口鱲有 2673 張，平頷鱲有 2580 張。

圖 10、粗首馬口鱲的範例照片。

圖 11、長鰭馬口鱲的範例照片。

圖 12、平頷鱲的範例照片。

## 3.2 模型架構

圖 13 是本論文的模型架構圖，我們是先將訓練集的圖片都經過前置影像處理，把圖片進行去背（在 3.3.2 有詳細說明），減少不必要的資訊後，再利用邊緣偵測框取物件，得到每張圖片被截取出來的主體圖片，之後把主體圖片透過關聯網路進行訓練。在關聯網路裡，*f* 是提取各個主體圖片特徵的模塊，將 sample set 的每張特徵圖與 query set 的每張特徵圖都進行配對。我們把原本關聯網路用串接的方式改成以相減的方式計算，因為兩張圖很相似的話，代表它們在空間上的距離是較近的，相減之後，它們之間的特徵圖差值會較小，我們希望模型能直接學習差值的變化。有了差值之後，就將差值再經過 *g* 模塊計算，輸出差值的關聯分數（relation score），若關聯分數大於或等於 0.5，代表相似度很高，會以 1 表示，反之相似的程度很低的話則為 0。如圖 13。

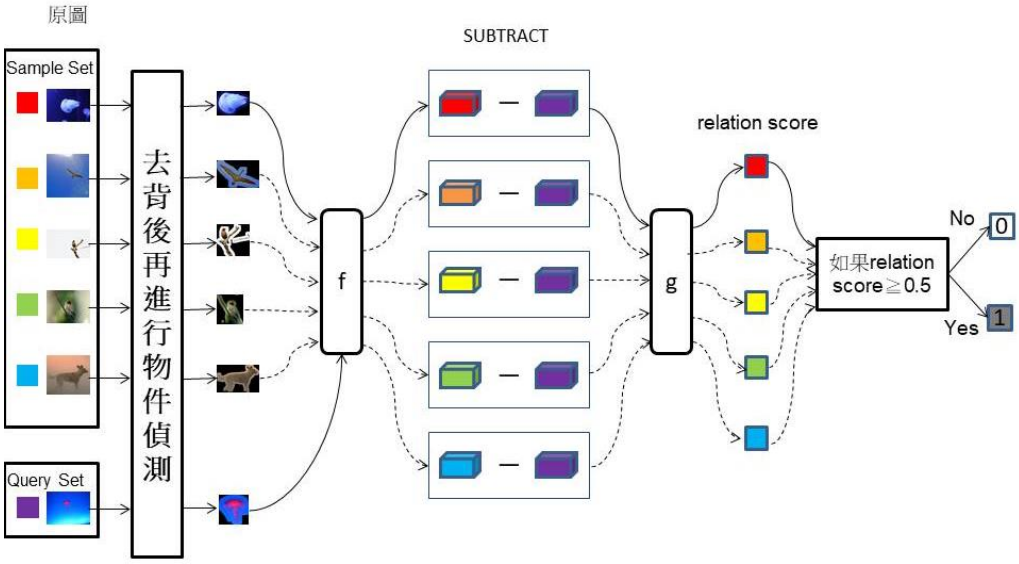


圖 13、本論文的模型架構圖。圖片來源為 miniImageNet。圖中的物件偵測是使用邊緣偵測。實線為當前正在計算相似度的配對，虛線為其它要計算相似度的配對。*f* 跟 *g* 是以 CNN 模型為基礎的網路架構，圖 14 有詳細的說明。以紅色代表跟紫色代表配對為例，兩者計算出的關聯分數高於 0.5，所以分類結果為塗上灰色的 1，表示這兩張照片是相似的。

關聯網路是藉由 *f* 模塊計算特徵圖，*g* 模塊計算相似度，其中 *f* 模塊是由 4 個 convolutional block 組成，前 2 個 convolutional block 各包含 1 個 2*2 的 max-pool，後 2 個不包含；*g* 模塊是由 2 個 convolutional block 各包含 1 個 2*2 的 max-pool 和

2 個 Fully Connected 層，一層為 ReLU，一層為 Sigmoid 作為輸出層組成。而每個 convolutional block 內包含一個 ReLU、一個 Batch Normalization、一個 3*3 卷積層及一個 64 filters。如圖 14。

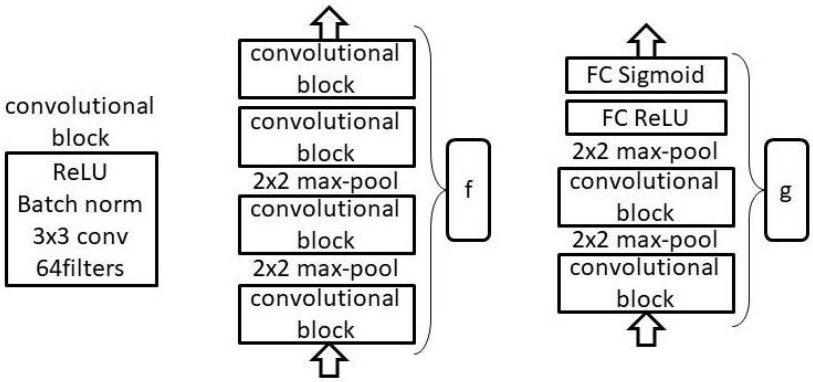


圖 14、網路結構組成圖。

在 2.2 節中有提到元學習會需要訓練兩個網路，learner和meta-learner。關聯網路的learner是卷積神經網路，meta-learner則是採用機器學習常見的optimizer。但是關聯網路採取的元學習法不適合用需要多次epoch才能收斂的optimizer，因為epoch主要目的是讓模型重複學習同樣資料，以此提高學習效果，但在元學習法中這反而會造成過度學習，然後導致模型過擬合的情況發生。所以關聯網路是選擇使用Adam[30]來做為optimizer的，因為Adam是不用經過多次迭代就可以快速收斂的optimizer。

## 3.3 物件偵測

### 3.3.1 僅選擇性搜索

我們直接使用OpenCV（https://opencv.org）內的選擇性搜索功能，該功能是被放在OpenCV Contrib中，所以要再自行額外安裝Contrib函式庫（https://github.com/opencv/opencv_contrib）。選擇性搜索有提供兩種模式，快速模式及精準模式。快速模式搜索的速度快，但找到的區域很少，所以精準度較差。精準模式則搜索很慢，但可以找到很多區域，精準度較佳。

我們是採用選擇性搜索的精準模式，將原圖直接進行區域搜索。由於精準模

式尋找到的區域數量很多，所以我們是假設照片主體是照片中最大的物件，因此會從尋找到的區域中，挑選面積最大的做為照片主體。但是選擇性搜索是依靠圖片分割後的色塊來做判斷，有可能會出現背景的區域面積比我們理想區域的面積大，所以如果直接讓原圖經過選擇性搜索，會有很高機率不會選擇到我們希望的區域，如圖 15。

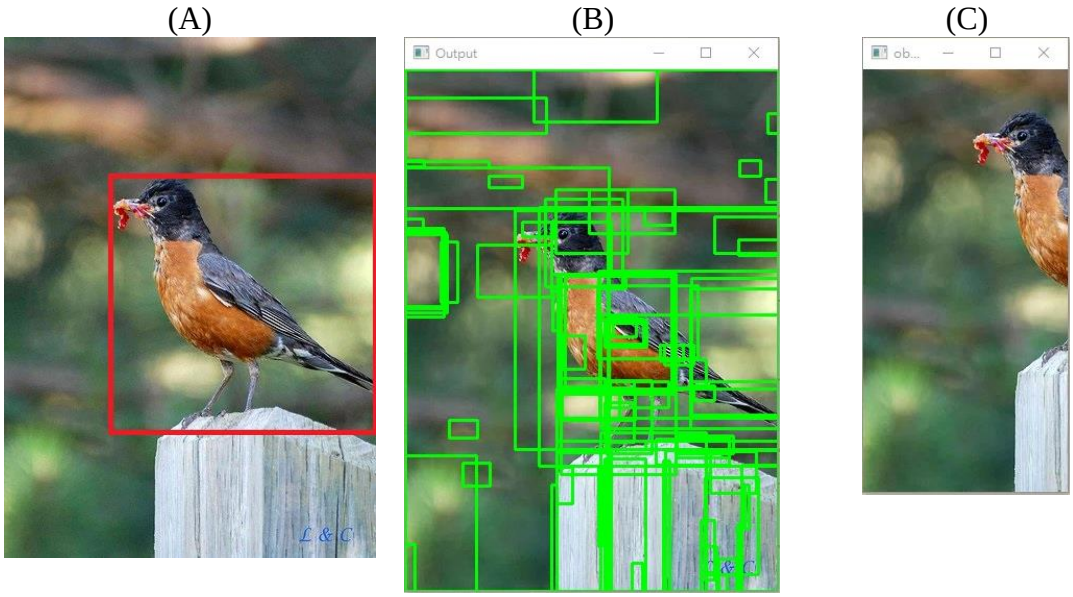


圖 15、將原圖直接進行選擇性搜索。(A)是原圖，我們理想的區域是紅框的範圍。(B)綠框是原圖透過選擇性搜索，尋找到的所有區域。(C)從找到的所有區域中，挑出面積最大的區域。會發現最後截取出的區域大多都是背景資訊，而鳥只被框進前半身。原圖來源 miniImageNet。

### 3.3.2 去背後，再進行選擇性搜索

由於僅靠選擇性搜索會無法準確截取到主體，如圖 15，所以我們在做物件定位前，如果先將圖片進行影像去背的處理，可以把不必要的資訊去除，一來能提高物件定位的準確度，二來去除了背景，主體被凸顯了出來。我們是採用自然影像去背，這方法需要Trimap來製作遮罩圖，但Trimap需要人工做前置處理，不符合我們的需求，所以我們用邊緣偵測製作了簡易的遮罩圖。OpenCV（https://opencv.org）提供了三種邊緣偵測的算法：Laplacian、Sobel跟Canny，我們是採用Canny算法[22]，它是目前較常用的算法，且邊緣的標記較準確清楚。

邊緣偵測要做影像去背的話，要先取得物件輪廓，然後填充輪廓。邊緣偵測

是使用灰階影像，然後用高斯模糊演算法去除部分雜訊。Canny 是以雙閾值來判定像素的灰階值是否保留成邊界，只要灰階值小於低閾值的不會被保留，大於高閾值的直接被保留，而介於之間的數值，若相鄰的灰階值有兩點是大於高閾值的，才會被保留。藉由像素的灰階值不同取得的邊界還並不是物件的輪廓，而是一連串沒有間斷的邊界才稱為輪廓，所以需經過適當的膨脹與腐蝕，膨脹是把亮的部分侵蝕暗的部分，腐蝕則是暗的部分侵蝕亮的部分，讓未相連的鄰近邊界連接起來，即可得到物件輪廓。之後將輪廓圈起的範圍以白色填充，就完成了該圖的遮罩圖。之後把遮罩圖跟原圖做 AND 計算，黑色部分在原圖上會被去除，白色部分會保持原樣，最後就得到一張去背後的圖。

圖像經過去背後，再進行選擇性搜索，會發現我們希望的部分，更容易被完整的框選進去了。而且去除了不相關的背景資訊，主體特徵會更加明顯，如圖 16。

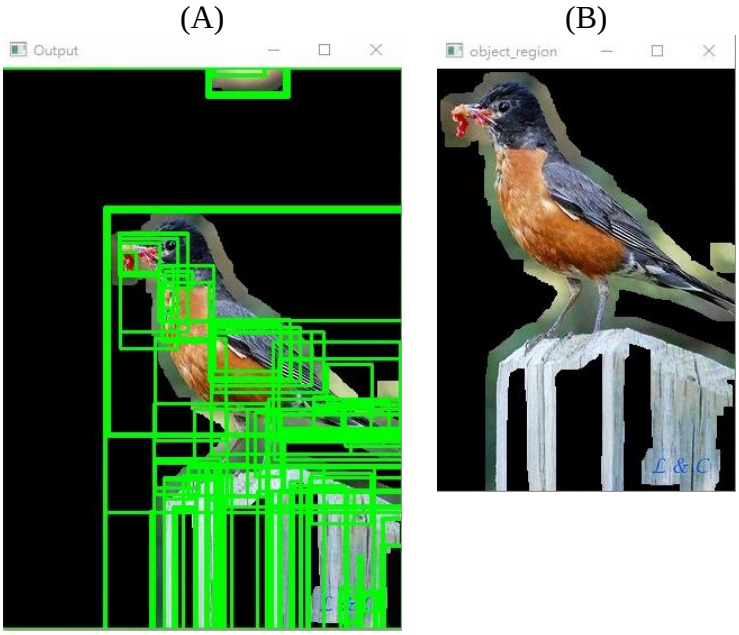


圖 16、將原圖去背後，再進行選擇性搜索。(A)去背後再經過選擇性搜索尋找到的所有區域。(B)裁切出在(A)中面積最大的區域。以此步驟來做，框到主體的準確度有提升。原圖來源 miniImageNet。

### 3.3.3 去背後，再進行邊緣偵測

除了選擇性搜索，我們還嘗試用邊緣偵測做物件的框選。因為我們發現經過去背後的選擇性搜索，不管是快速模式還是精準模式，還是會有機率出現框取不佳的結果，如圖 17(D)。

(A) (B) (C) (D)

圖 17、經過去背後的選擇性搜索，偶爾會出現不佳的框取。 (A)去背後，選擇性搜索尋找到的所有區域。(B)裁切出在(A)中面積最大的區域。(C)與(A)同張圖，但去背後，選擇性搜索尋找到與(A)略有不同的區域。(D)裁切出在(C)中面積最大的區域。

同 3.3.2 的去背方法，得到去背後的圖，把選擇性搜索改成邊緣偵測，因為被去掉的背景部分為黑色，與主體間的界線更為明顯，更容易抓到我們希望的範圍，如圖 18。以邊緣偵測來做框選物件不會產生只框取部分主體的現象，而且找最大面積也較快，因為選擇性搜索就算去背後還是會尋找到很多的候選框，如圖 17，回傳的數量可達 1813 個。而邊緣偵測回傳的區域數量會明顯的減少，數量只有 3 個，如圖 18。

(A) (B)

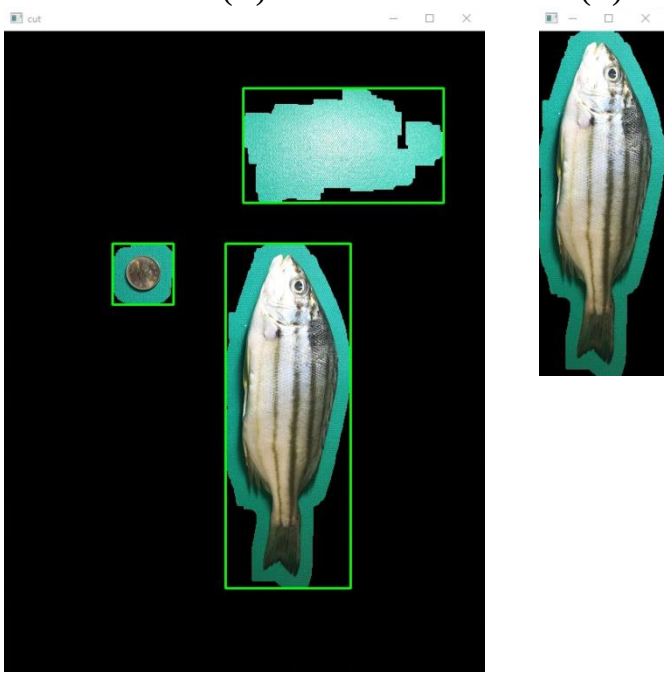

圖 18、去背後再進行邊緣偵測。(A)去背後用邊緣偵測索尋找到的所有區域。(B)裁切出在(A)中面積最大的區域。

# 四、實驗結果

## 4.1 Omniglot 的實驗結果

用 Omniglot 數據集訓練的 episode 設置為 10000 次，epoch 設置為 5 次。我們以 5-way 1-shot 包含 15 個 query 圖像進行訓練，也就是說每一個 episode 中會有 5*1+5*15 = 80 張圖像讓模型學習。

Omniglot 實驗上我們有兩個模型，一個是訓練集圖像是原圖的模型，這個模型會做為我們判斷的基準線；另一個是訓練集圖像有經過去背及邊緣偵測處理的模型。測試集的圖也會分成兩種，原圖跟經過去背及邊緣偵測處理的圖。

表 1、Omniglot 的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 原圖 | 0.987 |
| 去背後經過邊緣偵測 | | 0.977 |
| 原圖 | 去背後經過邊緣偵測 | 1.0 |
| 去背後經過邊緣偵測 | | 0.997 |

從表 1 的結果來看，可以發現 omniglot 不管是數據集的圖片還是測試集的圖片，有沒有經過去背及邊緣偵測處理，都對分類的準確度沒有多大的幫助。因為 omniglot 數據集的圖本身就很單純，畫面上只有字符，特徵本來就很明確，所以就算對 omniglot 數據集進行去背及邊緣偵測處理，訓練的效果跟以原圖進行訓練的效果是差不多。

## 4.2 MiniImageNet 的實驗結果

用 miniImageNet 數據集訓練的 episode 設置為 10000 次，epoch 設置為 5 次。我們以 5-way 1-shot 包含 19 個 query 圖像進行訓練，也就是說每一個 episode 中會有 5*1+5*19 = 100 張圖像讓模型學習。

MiniImageNet 實驗上我們有三個模型，一個是訓練集圖像是原圖的模型，這個模型會做為我們判斷的基準線；第二個是訓練集圖像有經過去背及選擇性搜索

處理的模型；第三個是訓練集圖像有經過去背及邊緣偵測處理的模型。測試集的圖也會分成三種，原圖、經過去背及選擇性搜索處理的圖跟經過去背及邊緣偵測處理的圖。

在 3.3.1 有提到自然圖像僅利用選擇性搜索處理的效果並沒有很好，會有部分圖像是沒有框在我們理想的區域上，反而是框到與主體不相關的背景，如圖 15(C)。因為影像處的效果不符合我們的需求，所以僅透過選擇性搜索處理後的圖，並未再做後續的影像分類訓練。

表 2、miniImageNet 三個模型對測試圖像是原圖的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 原圖 | 0.437 |
| 去背後經過選擇性搜索 | | 0.22 |
| 去背後經過邊緣偵測 | | 0.5 |

表 2 是我們的三個 miniImageNet 模型，分別對測試集圖像是原圖來進行預測分類。訓練圖像是以原圖進行訓練的模型，預測正確分類的準確度為 0.437，我們會以此模型的準確度做為表 2 的基準線。訓練圖像是有經過去背及選擇性搜索處理後進行訓練的模型，它的預測正確分類的準確度為 0.22。訓練圖像是有經過去背及邊緣偵測處理後進行訓練的模型，它的預測正確分類的準確度為 0.5。

從表 2 中可以看到經過去背及選擇性搜索處理的模型，它的準確度會比基準線低，原因是因為去背後再經過選擇性搜索有機率出現不完整或不正確的主體選取，如圖 19，導致模型在學習上是學習到不完整的特徵，或是學習到不是我們理想區域的特徵。而經過去背及邊緣偵測處理的模型，它的準確度比基準線高，以這方法進行處理比較不會出現不完整的主體選取，能讓模型專注學習到主體特徵，所以在原圖分類正確的準確度是提升的。

表 3、miniImageNet 三個模型對測試圖像經過去背及選擇性搜索處理的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 去背後經過選擇性搜索 | 0.427 |
| 去背後經過選擇性搜索 | | 0.18 |
| 去背後經過邊緣偵測 | | 0.413 |

表 3 也是我們的三個 miniImageNet 模型，但測試集圖像是有經過去背及選擇性搜索進行處理。訓練圖像以原圖進行訓練的模型，預測正確分類的準確度為 0.427，會以此準確度做為表 3 的基準線。訓練圖像是有經過去背及選擇性搜索處理後進行訓練的模型，預測正確分類的準確度為 0.18。訓練圖像是有經過去背及邊緣偵測處理後進行訓練的模型，預測正確分類的準確度為 0.413。

我們預想的情況是圖片經過去背及邊緣偵測訓練的模型，在分類圖片上準確度要比以原圖進行訓練的模型的準確度高， 表 2 結果是如同我們的預想準確度有提高，但在表 3 卻是下降的，這是因為測試集的圖片經過去背後經過選擇性搜索，有機率出現不完整或不正確的主體選取，如圖 19，這就是導致準確度降低的原因。

(A) (B) (C)

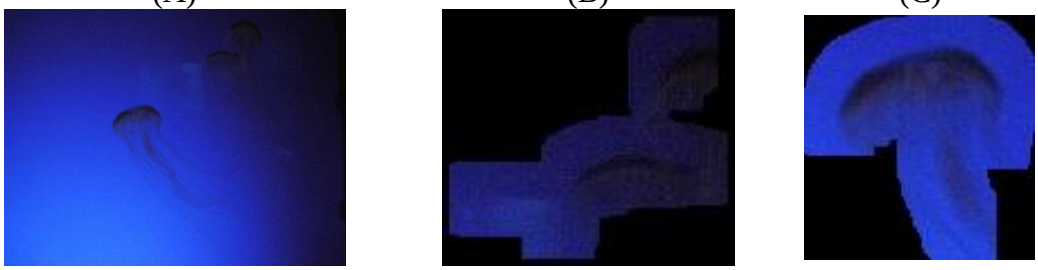

圖 19、去背後經過選擇性搜索產生的不理想照片，但去背後經過邊緣偵測產生較佳的效果。(A)原圖。(B)原圖去背後，選擇性搜索尋找到的區域。(C)原圖去背後，邊緣偵測尋找到的區域。原圖來源 miniImageNet。

(A) (B)

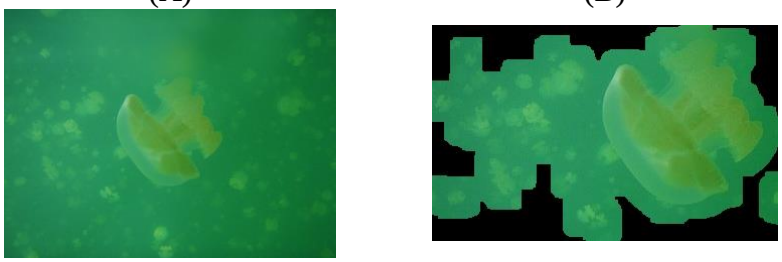

圖 20、去背後經過邊緣偵測產生的不理想照片。(A)原圖。(B)原圖去背後，邊緣偵測後尋找到的區域。原圖來源 miniImageNet。

表 4、miniImageNet 三個模型對測試圖像經過去背及邊緣偵測處理的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 去背後經過邊緣偵測 | 0.393 |
| 去背後經過選擇性搜索 | | 0.177 |
| 去背後經過邊緣偵測 | | 0.477 |

表 4 也是我們的三個 miniImageNet 模型，但測試集圖像是有經過去背及邊緣偵測進行處理。訓練圖像以原圖進行訓練的模型，預測正確分類的準確度為 0.393，以此準確度做為表 4 的基準線。訓練圖像是有經過去背及選擇性搜索處理後進行訓練的模型，預測正確分類的準確度為 0.177。訓練圖像是有經過去背及邊緣偵測處理後進行訓練的模型，預測正確分類的準確度為 0.477。

表 4 結果同我們預想的一樣，訓練圖像經過去背及邊緣偵測訓練的模型，它的準確度也是比基準線高。而訓練圖像是經過去背及選擇性搜索的模型，一樣是因為它不完整或不正確的主體選取的關係，導致準確度比基準線低。但也可以發現表 4 的三個準確度都比表 2 還低，原因是雖然邊緣偵測不會如同選擇性搜索出現主體被截取一半的情況發生，但還是有部分圖片物件選取並沒有很理想，如圖 20，所以導致分類的準確度會有所下降。

從這三個表的結果來說，圖片有經過去背及邊緣偵測處理後，的確是可以提高分類準確度的，雖然在表 3 中，經過去背及邊緣偵測訓練的模型的表現沒有比原圖訓練的模型好，但其實準確度並沒有差得太多，表示圖片經過去背及邊緣偵測處理後，再進行訓練的模型還是有不錯的分類效果。

我們有做原本採用串接方式的關聯網路的模型，以訓練集圖像是原圖時進行訓練，之後對測試集圖像是原圖來進行預測分類，它的準確度是 0.453。我們修改成相減計算的關聯網路，以相同條件進行訓練並預測分類，準確度是 0.437，比原本串接方式的關聯網路低。這是因為原本的關聯網路是採用可學習的非線性度量，而一般常用的距離度量，例如 L1 距離、歐式距離或 Cosine 相似度等，是固定的線性度量。這在有些數據看起來不相關，但在某些情況下可能是屬於同一類時，以固定的線性度量方式會學習不佳，而可學習的非線性度量則可以學習的不錯。

## 4.3 新國民食魚教育計畫的實驗結果

我們的新國民食魚教育計畫數據集和溪哥數據集，照片量跟 omniglot 和 miniImageNet 相比起來還要更少，如果直接延用 omniglot 和 miniImageNet 的訓練參數，會讓模型過度學習。所以我們兩個魚數據集的訓練參數，有根據數據集的照片量狀況分別做調整。

用新國民食魚教育計畫數據集訓練的 episode 設置為 1000 次，epoch 設置為 1 次。我們以 2-way 1-shot 包含 10 個 query 圖像進行訓練，也就是說一個 episode 中會有 2*1+2*10 = 22 張圖像讓模型學習。

新國民食魚教育計畫實驗上我們有兩個模型，一個是訓練集圖像是原圖的模型，這個模型會做為我們判斷的基準線；另一個是訓練集圖像有經過去背及邊緣偵測處理的模型。測試集的圖也會分成兩種，原圖跟經過去背及邊緣偵測處理的圖。

表 5、新國民食魚教育計畫的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 原圖 | 0.672 |
| 去背後經過邊緣偵測 | | 0.53 |
| 原圖 | 去背後經過邊緣偵測 | 0.514 |
| 去背後經過邊緣偵測 | | 0.918 |

因為新國民食魚教育計畫的圖片量不多，所以訓練集圖像也會充當為測試集的圖像。從表 5 來看可以發現說用原圖進行訓練的模型，在進行預測分類自己學習過的圖時，準確度只在 0.672，但用去背及邊緣偵測處理後進行訓練的模型，在進行預測分類自己學習過的圖時，準確度達到 0.918。而且去背後經過邊緣偵測的模型在分類原圖時，準確度有 0.53，但原圖訓練的模型在分類有經過去背及邊緣偵測的圖時，準確度 0.514，是較低的。這是因為新國民食魚教育計畫的照片背景單純，在經過去背及邊緣偵測處理時，可以準確的截取出魚的部分，模型能專注學習魚的特徵，因此會有較好的分類效果。

## 4.4 溪哥的實驗結果

用溪哥數據集訓練的 episode 設置為 3500 次，epoch 設置為 1 次。我們以 3-way 1-shot 包含 15 個 query 圖像進行訓練，也就是說每一個 episode 中會有 3*1+3*15 = 48 張圖像讓模型學習。

溪哥實驗上我們有兩個模型，一個是訓練集圖像是原圖的模型，這個模型會做為我們判斷的基準線；另一個是訓練集圖像有經過去背及邊緣偵測處理的模型。測試集的圖也會分成兩種，原圖跟經過去背及邊緣偵測處理的圖。

表 6、溪哥的實驗結果。

| 訓練集圖像 | 測試集圖像 | 準確度 |
|---|---|---|
| 原圖 | 原圖 | 0.73 |
| 去背後經過邊緣偵測 | | 0.616 |
| 原圖 | 去背後經過邊緣偵測 | 0.611 |
| 去背後經過邊緣偵測 | | 0.701 |

溪哥也同新國民食魚教育計畫一樣，因為圖片量少的關係，訓練集圖像也會充當為測試集的圖像。不過結果並沒有同新國民食魚教育計畫一樣，去背後經過邊緣偵測模型在分類自己學過的圖片時，準確度是 0.701，是比原圖模型在分類自己學過的圖片的準確度 0.73 低的。這是因為溪哥的圖經過去背後及邊緣偵測處理，有出現不正確的物件選取關係，所以準確度是比原圖模型低的。但是兩個模型在對另一方學過的圖進行分類時，準確度卻是相差不多，都介於 0.61 附近。

我們除了將原本關聯網路的串接計算改成相減計算之外，也嘗試過差值放大，利用差值平方及差值三方，分別來進行模型訓練，不過兩個差值放大的模型，效果都沒有比差值計算的效果來的好。原因有可能是原本不相似的照片，相減後距離已經很遠，再差值放大把距離拉更遠，導致模型無法收斂；或是資料量不足，差值放大的模型無法有效學習。

# 五、結論與未來方向

我們將影像去背及物件偵測加入關聯網路，來進行影像分類。其中影像去背及物件偵測的目的是要處理自然影像，讓它能如同白底黑字的字符數據集一樣，僅有主體在畫面上，以此讓模型更專注在學習主體的特徵。從實驗結果來看，自然影像有經過前置處理後，分類的準確度是有提升。

我們的影像處理方法還有改善的空間，因為我們只是簡略的處理圖片，主體周圍還是留有背景資訊。而且不管是選擇性搜索還是邊緣偵測，物件定位的效果都沒有特別精準。還有我們是假設主體是照片中最大的物件，但實際上不見得為最大，以圖 15 為例，程式最後挑選出的最大物件是除了鳥之外還包含牠底下站的木頭，但我們希望被選到的區域是只有鳥而已。內文中提到的 Trimap 方法，能解決去背不完整及物件定位不準確的問題，因為 Trimap 的去背效果較精細，不會讓主體附近還帶有背景，之後的物件偵測也會較為準確，但 Trimap 方法的缺點就是太耗費人力。

可以嘗試從邊緣偵測做出的遮罩圖中，將擬合前景物件的輪廓線取一個範圍，做為背景與前景交界處的未知區域，可產生類似 Trimap 的圖，但時間不足的關係我們沒有做出這項方法，未來可接續嘗試。

也可以考慮使用 yolo 的框取物件方法，yolo 的框取方法是採用 faster R-CNN 的 anchor box，但是 faster R-CNN 是設置固定的 anchor box 來做搜尋，而 yolo 則是利用 k-means 把數據集裡的原本就存有的所有物件框做分群，以產生它需要的 anchor box。所以如果是要用自己的數據集的話，就要自己畫製數據集的物件框。

在網路的選擇部分，可以考慮使用 LSTM 來做。匹配網路的作者有比較卷積神經網路及 LSTM 網路是否有差別，他們測試的結果是，在 omniglot 數據集上兩種網路沒有差別，但在 miniImageNet 數據集上，LSTM 的效果比卷積神經網路還要好。不過因為我們考量到我們的兩個魚數據集比 omniglot 跟 miniImageNet 較小的關係，所以網路還是選擇用卷積神經網路來做。

在少樣本學習問題中，除了以少量樣本訓練的少樣本學習外，還有零樣本學習的研究方向。它與少樣本學習不同，它是沒有為 sample set 提供圖像的，而是以每個類別的語義描述替代圖像數據。但之後要計算與 query set 圖像數據之間相似度的方式，可沿用少樣本學習模型。若零樣本學習採用類似的影像處理方法，讓 query set 裡的圖像變為單純的圖像，或許能更加容易提取到與 sample set 的語意描述更加貼近的特徵。

# 參 考 文 獻


[1] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, " In *International Conference on Machine Learning*, Beijing, China, vol. 32, pp. 647-655, 2014.

[2] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition, " In *International Conference on Machine Learning*, Lille, France, 2015.

[3] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier, "Learning a synaptic learning rule, " In *International Joint Conference on Neural Networks*, Seattle, WA, USA, 1991.

[4] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching Networks for One Shot Learning, " In *Neural Information Processing Systems*, Barcelona, Spain, pp. 3637-3645, 2016.

[5] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical Networks for Few-shot Learning," In *Neural Information Processing Systems*, Long Beach Convention & Entertainment Center, CA, USA, pp. 4077-4087, 2017.

[6] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to Compare Relation Network for Few-Shot Learning," In *Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, 2018.

[7] Junhua Wang, Zijiang Zhu, Jianjun Li, and Junshan Li, "Attention Based Siamese Networks for Few-Shot Learning, " In *IEEE International Conference on Software Engineering and Service Science*, Beijing, China, 2018.

[8] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum, "Human-level concept learning through probabilistic program induction, " *American Association for the Advancement of Science*, vol. 350, Issue 6266, pp. 1332-1338, 2015.

[9] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap, "Meta-Learning with Memory-Augmented Neural Networks, " In *International Conference on Machine Learning*, New York City, NY, USA, vol. 48, pp. 1842-1850, 2016.

[10] Danilo J. Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra, "One-Shot Generalization in Deep Generative Models, " In *International Conference on Machine Learning*, New York City, NY, USA, vol. 48, pp. 1521-1529, 2016.

[11] Shruti Jadon, "An Overview of Deep Learning Architectures in Few-Shot Learning Domain, " In *Computer Vision and Pattern Recognition*[Online], Available: https://arxiv.org/abs/2008.06365, 2020.

[12] Suvarna Kadam, and Vinay G. Vaidya, "Review and Analysis of Zero, One and Few Shot Learning Approaches, " In *Advances in Intelligent Systems and Computing book series*, Janusz Kacprzyk, Ed. Cham: Springer, vol. 940, pp. 100-112, 2019.

[13] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, "A Simple Neural Attentive Meta-Learner," In *International Conference on Learning Representations*, Vancouver, Canada, 2018.

[14] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory, " *Neuralcomputation*, vol. 9, Issue 8, pp. 1735-1780, 1997.

[15] Sachin Ravi, and Hugo Larochelle, "Optimization as A Model for Few-shot Learning, " In *International Conference on Learning Representations*, Toulon, France, 2017.

[16] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, " In *Computer Vision and Pattern Recognition*, Las Vegas, America, 2016.

[17] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu, "Object Detection with Deep Learning: A Review, " *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, Issue. 11, pp. 3212-3232, 2019.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation, " In *Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, 2014.

[19] Ross Girshick, "Fast R-CNN, " In *IEEE International Conference on Computer Vision*, Santiago, Chile, 2015.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks, " In *Neural Information Processing Systems*, Montréal, Canada, 2015.

[21] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection, " In *Computer Vision and Pattern Recognition*[Online], Available: https://arxiv.org/abs/2004.10934, 2020.

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector, " In *European Conference on Computer Vision*, pp.21-37, 2016.

[23] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders, "Selective Search for Object Recognition, " *International Journal of Computer Vision*, vol. 104, pp. 154-171, 2013.

[24] Pedro F. Felzenszwalb, and Daniel P. Huttenlocher, "Efficient Graph-Based Image Segmentation, " *International Journal of Computer Vision*, vol. 59, pp.167-181, 2004.

[25] John Canny, "A Computational Approach To Edge Detection, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, Issue. 6, pp. 679-698, 1986.

[26] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman, "Real-Time High-Resolution Background Matting, " In *Computer Vision and Pattern Recognition*, Nashville, Tennessee, pp. 8762-8771, 2021.

[27] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang, "Deep Image Matting, " In *Computer Vision and Pattern Recognition*, Honolulu, Hawaii, pp. 2970-2979, 2017.

[28] Marco Forte, and François Pitié, "F, B, Alpha Matting, " In *European Conference on Computer Vision*[Online], Available: https://arxiv.org/abs/2003.07711, 2020.

[29] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, " In *International Conference on Machine Learning*, Sydney, Australia, 2017.

[30] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization, " In *International Conference on Learning Representations*, San Diego, CA, USA, 2015.