國立高雄科技大學

智慧商務系

碩士論文

文字與圖片共同學習的探討-
以 ImageNet 圖像分類任務為例

Exploring the Joint Learning of Text and Images: A Case Study on ImageNet Image Classification Task

研究生：　　羅宇君

指導教授：　龔千芬 博士

中華民國一一二年六月

文字與圖片共同學習的探討-以 ImageNet 圖像分類任務為例

Exploring the Joint Learning of Text and Images: A Case Study on ImageNet Image Classification Task

研究生：　　羅宇君

指導教授：　龔千芬 博士

國立高雄科技大學

智慧商務系

碩士論文

A Thesis
Presented to
Department of Intelligent Commerce
National Kaohsiung University of Science and Technology
in Partial Fulfillment of the Requirements
For the Degree of
Master of Science
In
Intelligent Commerce

June,2023
Kaohsiung, Taiwan, Republic of China

中 華 民 國 一 一 二 年 六 月

國立高雄科技大學 (燕巢校區) 研究所學位論文考試審定書

智慧商務系 碩士班

研究生 羅宇君 所提之論文

論文名稱(中文): 文字與圖片共同學習的探討- 以ImageNet圖像分類任務為例

論文名稱(英/日/德文): Exploring the Joint Learning of Text and Images: A Case Study on ImageNet Image Classification Task

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 盧成憲 簽章

委 員 盧成憲 _____

賴穎如 _____

[illegible] _____

龔千芳 _____

指導教授 龔千芳 簽章

系所主管 許瓊文 簽章

中華民國 112 年 6 月 13 日

保存期限：永久

國立高雄科技大學學位論文著作權歸屬協議書

論文名稱：文字與圖片共同學習的探討 以ImageNet圖像分類任務為例　**論文種類：**□博士論文　☑碩士論文

研究生：[illegible]宇君　**系所名稱：**智慧商務系

指導教授：[illegible]千芳

共同指導教授：

茲為保障著作人著作權益，並就論文著作權之歸屬及事後權利行使方式，包括論文應如何公開發表、發表時應如何標示著作人姓名、論文事後可作何種修改以及未來應如何授權他人利用等事項，碩、博士生與指導（含共同指導）教授依下列原則達成協議：

一、碩、博士生所撰寫之論文，如指導（或共同指導）教授僅為觀念之指導，並未參與內容表達之撰寫，依著作權法規定，學生為該論文之著作人，並於論文完成時，即享有該論文之著作權，指導教授無法於事後主張為共同著作人，亦不得共同掛名為著作人。（著作權法第 10 條之 1）

二、如指導（或共同指導）教授不僅為觀念的指導，且參與內容之表達而與學生共同完成論文，且各人之創作，不能分離利用者，則為共同著作，學生與指導教授為論文的共同著作人並共同享有著作權，此等共同著作著作權（包括著作財產權及著作人格權）的行使，即應取得碩、博士生與指導（或共同指導）教授之共同同意後，始得為之。（著作權法第 8 條、著作權法第 40 條之 1 第 1 項）

三、依上述原則，本論文之著作權歸屬：

□ 研究生單獨擁有。

☑ 研究生與指導教授共同擁有。

□ 研究生、指導教授及共同指導教授共同擁有。

研究生：[illegible]宇君　日期：112 年 8 月 1 日

指導教授：[illegible]千芳　日期：112 年 8 月 1 日

共同指導教授：＿＿＿＿＿＿　日期：　年　月　日

文字與圖片共同學習的探討-以 ImageNet 圖像分類任務為例

學生：羅宇君　　指導教授：龔千芬 博士

國立高雄科技大學智慧商務系碩士班

摘 要

本研究試圖透過轉移學習的概念，將文字中所學習的知識轉移至圖像分類的任務中，本研究使用 Word2vec 學習詞遷入向量，並且將詞遷入向量作為圖像分類任務之期望輸出，以融合圖片與文字的知識，提升圖像分類模型之績效。因此本研究將建構 MLP 模型並輸出 100 維詞遷入向量以證明是否以詞遷入向量作為訓練標籤能夠帶來模型績效之提升。本研究使用遷移學習並搭配 InceptionV3 預訓練模型，以及 ImageNet 圖片資料庫進行模型建構。本研究所使用之資料分為兩類一是來自於 ImageNet 中的 ISLVRC 子集 2017 年版的開放圖片資料集包含了 1000 個生物或物品類別，每個類別至少有 800 張已標注圖片以上。二是維基所提供的超大型英文語料庫，共有 20GB 的大小。其中包含了兩千萬篇以上的文章。預測分析上，將會為各類別計算詞嵌入向量，最後使用 MLP 搭配 InceptionV3 預測圖片輸入的詞嵌入向量，並以模型輸出與類別的向量距離來判別模型的準確率，最後透過靶心偏移演算法針對輸入向量作微調。經分析結果發現，未使用靶心偏移其 Top-1 準確率為 0.625580，使用靶心偏移其 Top-1 準確率則是 0.781440，兩者相差了 0.1558，因此證明使用詞遷入向量作為類別能夠針對類別微調進而提高準確率。

關鍵字：機器學習、圖像分類

Exploring the Joint Learning of Text and Images: A Case Study on ImageNet Image Classification Task

Student：LUO -YU-JUN Advisors: Dr. Chien-Feng Kung

Graduate Institute of Intelligent Commerce
National Kaohsiung University of Science and Technology

ABSTRACT

This study aims to transfer the knowledge learned from text to the task of image classification using the concept of transfer learning. Word2vec is employed to learn word embeddings, which are used as the expected output for the image classification task. This approach integrates knowledge from both images and text to enhance the performance of the image classification model. In this study, an MLP model is constructed to output 100-dimensional word embeddings to examine whether using word embeddings as training labels can improve model performance. Transfer learning is employed, along with the InceptionV3 pre-trained model and the ImageNet image database, for model construction. The data used in this study consists of two types: the first type is the open image dataset from the 2017 version of the ISLVRC subset in ImageNet, which contains 1000 biological or object categories with at least 800 labeled images per category. The second type is a large English corpus provided by Wikipedia, with a total size of 20GB, containing over 20 million articles. For prediction analysis, word embeddings are calculated for each category, and the MLP, in conjunction with the InceptionV3 model, is used to predict the word embeddings for the input images. The model's accuracy is determined by measuring the distance between the model's output and the category's embedding vector. Finally, the target center-offset algorithm is applied to fine-tune the input embeddings. The analysis results show that the Top-1 accuracy without target center-offset is 0.625580, while the Top-1 accuracy with target center-offset is 0.781440, with a difference of 0.1558. This demonstrates that using word embeddings as category labels can fine-tune the model for each category and improve accuracy.

Keywords: machine learning, Image classification

誌 謝

在碩士的這段期間，首先，要先感謝我的指導教授龔千芬老師，從大學時期的專題到碩士階段的論文，老師以耐心與細心教導我們，遭遇問題時也會以各種不同的面相給予建議與幫助。從起初大學階段的人工智慧培訓到引流學長姐，教師幫忙講解。以及提供優秀的環境供我們專注地進行實驗。感謝老師這六年來給予我們的幫助。

而第二要感謝的是郝沛毅教授，起初的論文發想是透過與老師的討論而開始有了模樣，並在論文的製作過程中指引及建議我們可前進的方向。在每周的開會討論上對於我們的問題提出更深度的見解進而幫助我們有了解決的思路。非常兩位感謝老師在百忙之中能夠幫助我完成此篇成就。

接著我想感謝我的同學與朋友，在碩士路上一路扶持，互相分享經驗與一齊討論問題，教學相長的方式讓我發現問題進而改進。

最後我想感謝在這段期間，給予我幫助的家人們，有你們的扶持才可以讓我順利完成這段碩士學位。

對於給我過幫助的所有人，有了你們才可以讓我圓滿完成我的碩士生涯

羅宇君 謹誌

中華民國 一一二年六月

目錄

表目錄

圖目錄

壹、緒論

1.1 研究背景與動機

自從 alpha Go 擊敗人類圍棋棋王李世乭，開啟了新一波的人工智慧熱潮，各大企業紛紛投入大量資金開發人工智慧系統，例如 OpenAI 開發的對話模型 GPT、藉由描述圖片效果產出對應圖片的 DALL.E 與語音辨識模型 Whisper，Google 最新推出的生成式模型 Bard，凸顯了人工智慧在現今的流行性。

面對如此狂熱的 AI 風潮，對於熟悉人工智慧發展史而言，不經疑問這次風潮會持續多久?更重要的是，掀起這次人工智慧高潮的深度學習，即是過去曾被為人詬病的聯結主義的延伸(Garson et al., 1997)。雖然現今有如 RELU、Dropout、K-Fold 等演算法來優化梯度下降(Vinod et al., 2010)、Over-fitting(Nitish et al., 2014)、樣本不足(Mosteller F. , et al. 1968)等問題。但本質上還是依靠反向傳播法來學習(David et al., 1986)、以及需要大量的人工標記的訓練資料來訓練模型。

對於人工智慧而言，需要大量的人工標記資料才可能建構的模型對整體的發展受到了非常大的限制。但如果沒有標記訓練資料的類別，類神經網路是無法學習、辨識的。雖然有許多的研究被提出，緩解建構大量訓練資料的昂貴成本、比如半監督式、弱監督式、遷移學習、生成對抗網路等等。

且目前的類神經網路結構還有一個限制，即為一個輸出節點對應一個類別，每一個類別都視為彼此獨立的，而不論類別與類別之間的關聯性，但是，「貓」與「老虎」這兩個類別同為貓科的關聯性，明顯是大於「貓」與「腳踏車」這兩個關聯性較弱的類別，如果能在訓練之前便知道每個類別的之間的關聯性，而不是一味地把每個類別都視為獨立，用不同的輸出節點來代表，以此能更有效率的捕抓到關鍵的特徵。

但現今的人工智慧技術，圖片與文字被視為兩大互不相干的領域，以文字探勘的領域中的 Word2vec 為例，它採用非監督的演算法，透過由前後文來預測當前文，或是由當前文來預測前後文，來學習字詞的語意相似性，並建立詞嵌入向量(word embedding vector)，語意越相近的兩個詞，其對應的詞嵌入向量之間的距離也會越小。在圖片辨識的領域中，眾所皆知最熱門的技術是卷積神經網路(CNN)，它是以監督式的學習策略，藉由對圖片進行卷積處理與標籤標記進行學習，而它標記的類別，不論是狗還是車子，也都是文字。若能將圖片的類別名稱從非監督式學習得到的結果，導入圖片的監督式學習演算法，巧妙融合監督式與非監督式、圖片跟文字之間的特性，可以分別克服彼此的缺點，創造更優異的學習成效。

綜上所述，為了讓這波人工智慧的熱潮可以繼續延續，我們需要開發嶄新的學習架構，來解決上述問題。

1.2 研究目的

本研究將建構 MLP 模型並輸出 100 維詞遷入向量，並藉由各列別詞遷入向量與輸出向量之間的距離計算出類別判別率，以證明是否以詞遷入向量作為訓練標籤能夠帶來模型績效之提升。本研究使用遷移學習並搭配 InceptionV3 預訓練模型，以及 ImageNet 圖片資料庫進行模型建構，並統整上述內容為以下三點:

1. 將 ImageNet 圖片資料搭配 InceptionV3 預訓練模型進行遷移學習，並提取輸出前的特徵向量，進行 MLP 模型訓練。

2. 使用 Gensim 開源 Word2vec 模型並搭配 wiki 開源語料庫建構各類別之詞遷入向量，進行 MLP 模型訓練。

3. 將 MLP 模型輸出結果與各類別之詞遷入向量做比對，觀察是否以詞遷入向量作為訓練標籤能夠帶來模型績效之提升。

1.3 論文架構:

本論文共包含五個章節,內容簡述如下:

第一章 緒論

主要介紹目前深度學習之現況，並帶出圖片與文字共同學習之理念，最後說明本研究之主要研究目的

第二章 文獻探討

說明本研究所使用之演算法相關文獻，以及與本研究相似之模型架構與其成效。

第三章 研究方法

包含本研究之研究架構、資料來源、資料說明、介紹本研究使用之機器學習方法並詳細說明處理流程，包括資料預處理、模型訓練、模型輸出及後續處理。

第四章 研究結果

將實際實行之結果透過圖表及文字進行解釋，並對靶心偏移演算法進行說明、檢討其效用。

第五章 結論與建議

總結本研究之研究結果、貢獻，說明本研究之限制以及缺失並給予未來研究者相關之研究建議。

貳、文獻探討

2.1 自然語言處理

自然語言處理(NLP)是一門由語言學、電腦科學與人工智慧組成的領域。NLP 致力於收集關於人類如何理解與使用語言的知識，並以此為基礎開發適當的工具與技術，使電腦系統性的理解和使用自然語言。特別關注於電腦與自然語言之間的互動，特別是如何為編程已處理與分析大量的自然語言數據。NLP 的應用包含許多研究領域，如機器翻譯、自然語言文本處理、用戶介面、多語言和跨語言搜尋系統、語音辨識、人工智慧、專家系統等等。

2.1.1 Word2vec

Word2vec 是於 2013 年發表的一項自然語言處理（NLP）技術。Word2vec 使用神經網路模型從大型文本與料庫學習單字的關聯性。經過訓練的模型可以判斷出同義詞或部分與部分句子有關聯的字詞。Word2vec 使用的是稱為向量(vector)的特定數字陣列來表示每個不同的單字。這些向量會因為字詞之間的關聯性而有所不同，可以藉由數學函數計算後的距離來表明向量所代表的單詞之間語意相似程度。Word2vec 可以使用連續詞袋（CBOW）或連續跳格(skip-gram)這兩種架構來表示字詞向量。在這兩種架構中，Word2vec 在對整個語料庫進行訓練時，都會考慮字詞與字詞之間周圍的上下文詞的滑動窗口(Sliding Window)。在連續詞袋架構中，該模型會從周圍的上下文詞的窗口中預測當前的詞。語意詞的順序並不影響預測。而在連續跳格架構中，模型使用當前詞來預測周圍的語意詞窗口。跳格結構比起較遠的語意詞，會對較近的語意詞更為重視。在模型訓練完成後，在語料庫中具有共同語境的詞，即在語義和句法上相似的詞彙在向量中彼此靠近。

2.2 機器學習

機器學習是一個已建立”學習”方式的研究領域。換言之，利用數據來提高某些任務的成功率。機器學習算法基於數據建立模型，並運用模型做出預測或決定。機器學習被廣泛運用在各種領域中，如醫學、電子學、語音辨識、農業、電腦視覺等等。

2.2.1 遷移學習

遷移學習是屬於人工智慧的一種學習領域，其重點在於如何使用與訓練目標部分相關的已訓練模型來輔助模型更好的訓練目標，進而達到減少訓練所需的資料量、加快訓練流程等等。遷移學習的學習方式可以分為以下三類: 推導遷移學習（inductive transfer learning），轉導遷移學習（tranductive transfer learning）和無監督遷移學習（unsupervised transfer learning）。

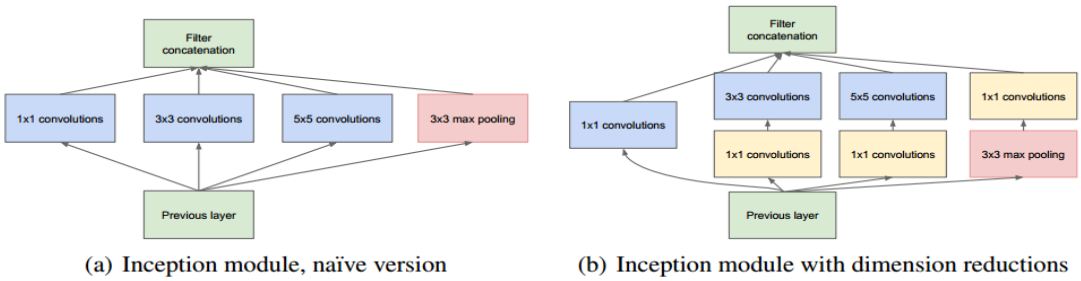
推導遷移學習:在原模型學習目標與訓練目標不同的情況下，使用原模型訓練結果與新訓練領域來提升或優化訓練目標的學習效果。

轉導遷移學習:與推導遷移學習不同的地方在於，轉導遷移學習除了要求訓練目標的不同外、原模型的訓練領域也要與新增的訓練領域不同。且需要而外提供一些新訓練領域的無標記數據。

無監督遷移學習:與推倒遷移學習所需的條件相同，但差別在標籤不可觀測。無監督遷移學習主要解決的目標領域中的無監督學習問題。類似於降維、密度估計等機器學習問題。

2.2.2 InceptionV3

InceptionV3 是基於 InceptionV2 所改良出來的模型架構。如圖 1 所示，該系列架構最大的特點在於有特殊設計的 Inception Module。將不同 branch 中獲得的 feature map 拼接再一起，並採用 Same paddle 的方式讓每張 feature Map 相同。且為了降低計算量，導入了幾個 1*1 的卷基層來降低維度。



(a) Inception module, naïve version　　(b) Inception module with dimension reductions

圖 1　Inception Module 架構圖

資料來源: Christian,2014, Going Deeper with Convolutions

InceptionV3 與前作相比，它將固定的 3*3、5*5 卷積層轉換成 1*7、7*1 等不對稱卷積。與對稱卷積的成效相比，不對稱卷積不僅降低了參數量同時也增加了模型的深度。第二是發現了輔助分類器的做用，作者發現將輔助分類器剃除後不會對模型產生不良影響。此外若加入的如 Batch Normalization 或是 Dropout 層會使得模型效能更好，因此認為輔助分類器有正則化的作用。第三是從原本在 Same paddle 時的 stride 從 1 改為 2 來進一步的縮小特徵圖。且輸入圖像大小改為 299*299。

2.2.3 MLP

多層感知器(MLP)是一個全連接型的人工神經網路、一個 MLP 需要有三個結構(輸入層、隱藏層、輸出層)，其中會使用反向傳播法來進行模型訓練。反向傳播的目的就是利用最後的目標函數(loss/cost function)來進行參數更新，一般都會使用誤差均方和(MSE)當作目標函數，若 MSE 越高代表模型越差。需要進行進一步的學習直到參數或誤差值收斂。

2.2.4 SGD

隨機梯度下降法（通常縮寫為 SGD）是一種反覆運算方法，用於優化具有適當平滑特性（如可微分或次微分）的目標函數。它可以被視為梯度下降優化的隨機近似，因為它用其估計值（從隨機選擇的資料子集計算）取代了實際梯度（從整個資料集計算）。特別是在高維優化問題中，這減少了非常高的計算負擔，實現了更快的反覆運算，以換取較低的收斂率。與 GD 不同的是，SGD 每經過一次訓練就會更新一次參數，且每次訓練所參照的樣本或 mini-batch 是隨機的。

2.3 相關研究文獻

以下將簡述兩篇以圖文結合為基礎並以此創建出的模型架構。

在 2021 年 Wonjae Kim 等作者創造出使用 transformer 處理視覺特徵並在預訓練過程使用 whole word masking 和圖像增強技術已提高模型表現。作者認為，早期的圖文結合工作只著重在圖片特徵擷取與編碼上，而對於文本與圖像間的互動都還停留在淺層。因此作者提出的 ViLT 架構使文本與圖像可以在深層次進行互動的方式。

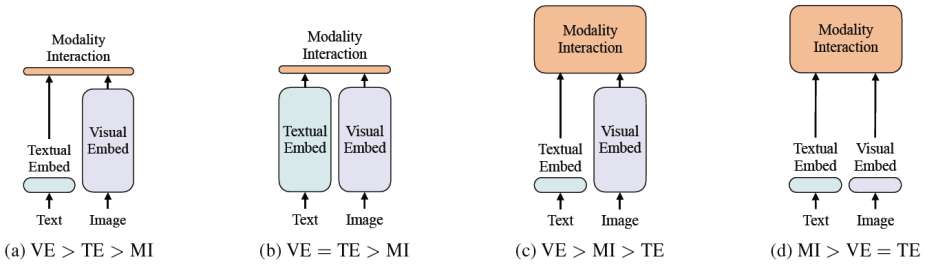


圖 2　ViLT 流程簡述圖

資料來源: Wonjae,2021, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

如圖 2 所示，流程圖(a)所代表的典型演算法為:VSE(2017)與 SCAN(2018);圖(b)代表的則是 CLIP(2021);圖(c)則是目前圖文結合模型選擇最多的架構;最後的圖(d)則是指 ViLT 模型。

圖 3 為 ViLT 模型架構圖，作者將文本資料與圖片資料分別用 0 與 1 代表，文字的預處理只有經過 word embedding 圖片則是一層卷積層做處理。並且將資料使用 transformer 讓圖片與文字互相交疊後進行訓練任務。

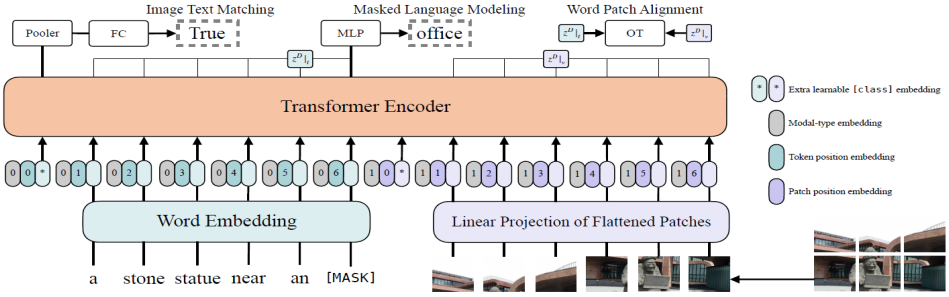


Figure 3. Model overview. Illustration inspired by Dosovitskiy et al. (2020).

圖 3　ViLT 流程簡述圖

資料來源: Wonjae,2021, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

成效部分作者將以圖 4 零成本學習與圖 5 一般任務分割開來，並區分文字檢索與圖片檢索。ViLT-B/32 為本篇作者之模型架構，其中表現較好的是 R@5 與 R@10。在三種資料集中準確率不輸給其他同類型模型。值得注意的地方在於運行時間，ViLT-B/32 所需的運行時間與其他同種類模型相比是非常的少，卻能獲得與類似架構相差無幾的準確率為 ViLT-B/32 模型的一大量點。

Visual Embed	Model	Time (ms)	Zero-Shot Text Retrieval						Zero-Shot Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	ViLBERT	~900	-	-	-	-	-	-	31.9	61.1	72.8	-	-	-
	Unicoder-VL	~925	64.3	85.8	92.3	-	-	-	48.4	76.0	85.2	-	-	-
	UNITER-Base	~900	80.7	95.7	98.0	-	-	-	66.2	88.4	92.9	-	-	-
	ImageBERT†	~925	70.7	90.2	94.0	44.0	71.2	80.4	54.3	79.6	87.5	32.3	59.0	70.2
Linear	ViLT-B/32	~15	69.7	91.0	96.0	53.4	80.7	88.8	51.3	79.9	87.9	37.3	67.4	79.0
	ViLT-B/32⊕	~15	73.2	93.6	96.5	56.5	82.6	89.6	55.0	82.5	89.8	40.4	70.0	81.1

圖 4　ViLT 零成本學習模型成效圖

資料來源: Wonjae,2021, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

Visual Embed	Model	Time (ms)	Text Retrieval						Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	w/o VLP SOTA	~900	67.4	90.3	95.8	50.4	82.2	90.0	48.6	77.7	85.2	38.6	69.3	80.4
	ViLBERT-Base	~920	-	-	-	-	-	-	58.2	84.9	91.5	-	-	-
	Unicoder-VL	~925	86.2	96.3	99.0	62.3	87.1	92.8	71.5	91.2	95.2	48.4	76.7	85.9
	UNITER-Base	~900	85.9	97.1	98.8	64.4	87.4	93.1	72.5	92.4	96.1	50.3	78.5	87.2
	OSCAR-Base†	~900	-	-	-	70.0	91.1	95.5	-	-	-	54.0	80.8	88.5
	VinVL-Base†‡	~650	-	-	-	74.6	92.6	96.3	-	-	-	58.1	83.2	90.1
Grid	Pixel-BERT-X152	~160	87.0	98.9	99.5	63.6	87.5	93.6	71.5	92.1	95.8	50.1	77.6	86.2
	Pixel-BERT-R50	~60	75.7	94.7	97.1	59.8	85.5	91.6	53.4	80.4	88.5	41.1	69.7	80.5
Linear	ViLT-B/32	~15	81.4	95.6	97.6	61.8	86.2	92.6	61.9	86.8	92.8	41.3	72.0	82.5
	ViLT-B/32ⓐ	~15	83.7	97.2	98.1	62.9	87.1	92.7	62.2	87.6	93.2	42.6	72.8	83.4
	ViLT-B/32ⓐ⊕	~15	83.5	96.7	98.6	61.5	86.3	92.7	64.4	88.7	93.8	42.7	72.9	83.1

圖 5　ViLT 一般任務學習模型成效圖

資料來源: Wonjae,2021, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

同為 2021 年的文章，Zhicheng Huang 等學者為了解決 ROI 只能獲取到一部分特徵的缺陷作者使用了一套 VD-base embedding(Visual Dictionary)模組來針對圖片進行處理。如圖 6 所示 SOHO 其中的特點在於不向 ROI 針對特定區域輸入，而是選擇使用整張圖片作為輸入，讓編碼器可以從預訓練 loss 或終端任務的 loss 中進行端對端的更新。作者提出了使用 VD-base embedding 模組，將與圖片相似的視覺語意合成到圖片特徵中，以獲得對應的 token。

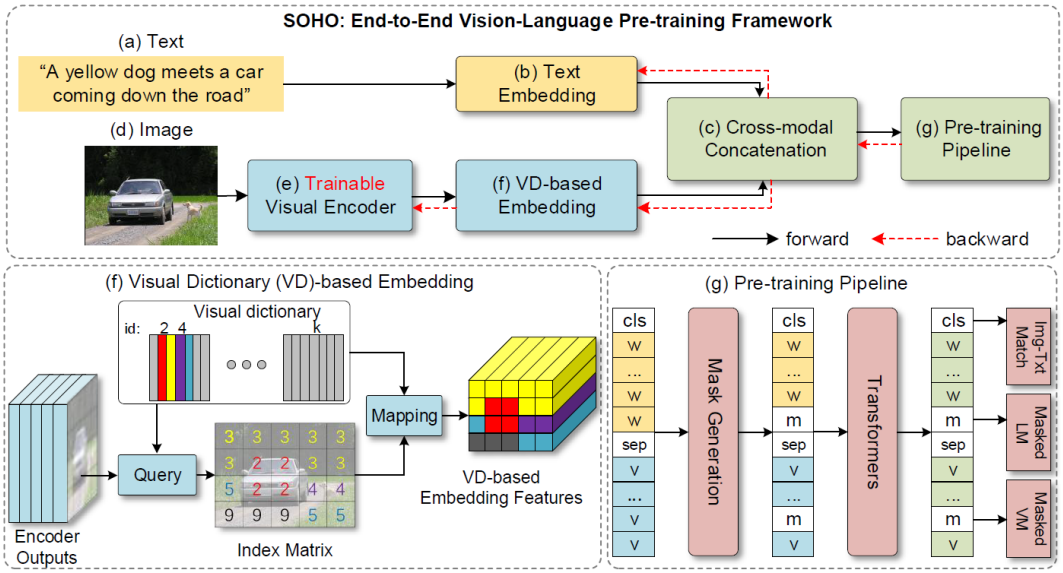


圖 6　SOHO 模型架構圖

資料來源: Zhicheng,2021, Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning

最後成效的部分，以其中的圖文、文圖預測結果為例。在 1K 測試集中與其它模型相比皆為 SOHO 最高。5K 測試集則在文圖預測中的 R@5、R@10 略輸給其他模型。

Model	Backbone	TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
		1K Test set						5K Test set					
VSE++[12]	R152	64.6	90.0	95.7	52.0	84.3	92.0	41.3	71.1	81.2	30.3	59.4	72.4
SCAN[21]	R101	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
Unicoder-VL[22]	-	84.3	97.3	99.3	69.7	93.5	97.2	62.3	87.1	92.8	46.7	76.0	85.3
UNITER[7]	R101	-	-	-	-	-	-	64.4	87.4	93.1	50.3	**78.5**	**87.2**
SOHO (ours)	R101	**85.1**	**97.4**	**99.4**	**73.5**	**94.5**	**97.5**	**66.4**	**88.2**	**93.8**	**50.6**	78.0	86.7

圖 7　SOHO 的圖文、文圖預測準確率

資料來源: Zhicheng,2021, Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning

參、研究方法

本研究使用之資料來自於 ImageNet 中的 ISLVRC 子集 2017 年版的開放圖片資料集，並使用自然語言模型、轉移學習等方式來驗證使用詞嵌入向量作為類別的期望輸出向量，是否會幫助圖片模型進行學習。本章節將說明整體研究架構、資料來源及預處理、模型架構及環境、模型訓練及績效評估方法。

3.1 研究架構

本研究使用之資料來自於 ImageNet 中的 ISLVRC 子集 2017 年版的開放圖片資料集，以及 wiki2023 年 2 月英文與料庫進行 Gensim 自然語言模型訓練，目的為計算各類別之詞嵌入向量，最後使用 MLP 搭配 InceptionV3 預測圖片輸入的詞嵌入向量，並以模型輸出與類別的向量距離來判別模型的準確率。目標為觀察使用詞嵌入向量作為類別的期望輸出向量，是否會幫助模型學習。

3.2 資料來源

本研究兩種開源資料庫，首先是 ImageNet 開放資料庫中的 ISLVRC 子集 2017 年版，該版本包含了 1000 個生物或物品類別，每種類別至少有 800 張已標注圖片以上。例如:「硬碟」或「獅子」等等。第二種是維基所提供的超大型英文語料庫，共有 20GB 的大小。其中包含了兩千萬篇以上的文章。

3.2.1 資料說明

ISLVRC 中的圖片分成訓練、測試、驗證資料集，每種資料集中皆有對應的圖片與標注檔，圖片對應的類別名稱存在著複數個單詞，且每種類別有用自己的類別編號。每張圖片的解析度大小不一，從 80*80 至 400*400 之間。ISLVRC 中的標注檔存在兩種注釋:(1)圖像類別的注釋，例如:”此圖像有腳踏車，但是沒有貓咪”(2) 圖像標注，即為將目標物件的周圍使用座標軸以方塊形狀的方式記錄起來。例如:”有一個鍵盤，為於(5,5)的位置，寬度 50，高度 25。

3.3 Word2vec

使用非監督式 Word2vec 學習物件名詞的詞嵌入向量，並且用詞嵌入向量當作圖像分類時的期望輸出向量，來巧妙地結合非監督與監督式學習的優點，並且將文字學習得到的知識(即詞嵌入向量)，透過轉移學習的概念來移轉給圖像分類的任務，以文字學習得到的知識來幫助圖像分類的任務，故我們使用 Word2vec 取得 ImageNet 的 1000 個類別的詞嵌入向量作為本論文的標籤

3.3.1 資料預處理

維基語料庫的資料處理，為了避免語料庫中擁有重複詞義的關鍵字，比如同樣代表金魚但擁有俗稱”goldfish” 與學名 “Carassius auratus”這兩種單字混淆訓練。我們將重複詞義的詞統一替換成相同單字，並加入各類別的首單字以避免維基資料庫不包含相關字詞而造成類別不完整。

3.3.2 模型訓練

將處理過後的維基語料庫轉換成 LineSentence 使資料可以迭代，並透過 Gensim 套件進行 Word2vec 模型訓練。其中，輸出向量大小設定成維度 100 維，前後詞的計算範圍設定為 10，並且將出現過 1 次以上的詞納入考量以避免標籤缺漏，進行 10 個 epochs 進行訓練。

3.3.3 Word2vec 結果輸出

訓練過後的 Word2vec 模型會進行標籤提取，在輸出時避免標籤與驗證資料錯位，會對標籤替換空格與底線做到規格統一，以及按照官方標籤順序輸出。輸出格式如圖 8 所示。

```
1.kit_fox：[ 0.08388428 -0.16280235 -0.25121623 -0.10876942  0.05156184  0.0033505
 -0.00654314 -0.02410968 -0.15018182  0.13763271 -0.10217406 -0.31039074
  0.23886105 -0.03907521 -0.14868715  0.01286481  0.1275539   0.3451754
 -0.25473422  0.10937279  0.20065641  0.03376088  0.52919096 -0.10169461
 -0.07195729 -0.37545973  0.34814838  0.46345574 -0.3319667   0.01840213
 -0.13013399 -0.15805198  0.19497623 -0.11544885  0.44164136 -0.1643137
  0.22695157  0.38761216  0.23787655  0.5472292  -0.18812665  0.405234
  0.4867272  -0.18972248 -0.1883036  -0.05220128  0.14115113 -0.11528226
 -0.07513387 -0.14349782 -0.24221286  0.06721443  0.19717966 -0.06096532
  0.46362603  0.07431096 -0.13760473 -0.19433652  0.05850148  0.09995046
 -0.03141404  0.15988338  0.02393502  0.06590812 -0.20169182 -0.03774843
  0.07379623 -0.04012696  0.05391183  0.3231753  -0.04954046 -0.3077464
  0.04758467 -0.2184537  -0.13675627  0.10221569 -0.2326045   0.00491163
 -0.09583104 -0.12689844  0.02767111  0.19961044  0.08936748  0.24843751
  0.35219613  0.08661895 -0.42494586 -0.07641659 -0.15051606 -0.13920492
  0.119836   -0.22305197 -0.17932     0.48615876 -0.34930822  0.33558238
 -0.1672039   0.14550385  0.08348957 -0.29685643 -0.00965938  0.31360537
  0.05430751  0.2747658  -0.15881658 -0.48586774 -0.11376674  0.2015545
 -0.12546259 -0.02379173  0.01690668  0.18973325 -0.04553036  0.17356272
  0.06443703 -0.46162906  0.11531527 -0.0555345  -0.0539911  -0.04093106]
2.english_setter：[-7.85437897e-02 -5.79834506e-02 -1.78032100e-01  2.69506127e-01
  4.15130965e-02  1.88520133e-01 -2.34186828e-01  2.69892186e-01
  1.67956382e-01  3.86665482e-03  1.03790186e-01  1.91955894e-01
  3.39615405e-01  1.77433223e-01  8.19285363e-02  6.90914169e-02
  1.15781710e-01  9.61361378e-02 -5.05209193e-02  7.21668005e-02
  1.73391014e-01  2.43881438e-02  7.12334886e-02 -3.26260507e-01
 -1.53159380e-01  1.35870829e-01 -8.67975652e-02  2.00158447e-01
 -3.32190879e-02 -2.87335098e-01  7.48866126e-02 -2.26811588e-01
  1.83102325e-01  3.82945165e-02 -1.90317556e-01  2.95367271e-01
  5.13953790e-02 -1.19995162e-01 -1.92393571e-01  2.89563179e-01
```

圖 8　Word2vec 詞遷入向量輸出

資料來源:本研究

3.4 InceptionV3

InceptionV3 作為圖片處理的預訓練模型，它擁有的 Inception Module 可以幫助我們的圖片參數量的降低以及特徵圖的縮小，同時不丟失大量的特徵訊息。並搭配 Global Average Pooling(GAP)將原圖片的參數量更進一步的降低，避免模型肥大化，故我們使用 InceptionV3 取得 ImageNet 的圖片特徵作為本論文的特徵。

3.4.1 資料預處理

在 ISLVRC 的標注檔中內含圖片標注框，為了避免特徵失焦，我們將圖片進行了標注框的擷取作為圖片輸入。並將圖片解析度統一至 299*299 對應原論文輸入，最後獲得輸入特徵(X)。為了能獲得與 X 的輸出標籤(Y)，我們將上一小節所獲得的 Word2vec 輸出與 X 做對應，具體之格式如圖 9 所示。X 代指輸入特徵，Xi_j,k 分別為 i:類別索引、j:資料索引、k:特徵索引。N 為 VGG 的特徵數量，代指 2048。X1,1 是代表第 1 個類別的第一張圖片之 InceptionV3 特徵。Y 代指期望輸出，Yi,h 分別為: i:類別索引、h:詞遷入向量之維度索引。M 為 Word2vec 模型之輸出向量，代指 100。

	輸入特徵					期望輸出			
第1個類別	第1筆資料	X1,1,1	X1,1,2	…	X1,1,N	Y1,1	Y1,2	…	Y1,M
	第2筆資料	X1,2,1	X1,2,2	…	X1,2,N	Y1,1	Y1,2	…	Y1,M
	第1000筆資料	X1,1000,1	X1,1000,2	…	X1,1000,N	Y1,1	Y1,2	…	Y1,M
第2個類別	第1筆資料	X2,1,1	X2,1,2	…	X2,1,N	Y2,1	Y2,2	…	Y2,M
	第2筆資料	X2,2,1	X2,2,2	…	X2,2,N	Y2,1	Y2,2	…	Y2,M
	第1000筆資料	X2,1000,1	X2,1000,2	…	X2,1000,N	Y2,1	Y2,2	…	Y2,M
第1000個類別	第1筆資料	X1000,1,1	X1000,1,2	…	X1000,1,N	Y1000,1	Y1000,2	…	Y1000,M
	第2筆資料	X1000,2,1	X1000,2,2	…	X1000,2,N	Y1000,1	Y1000,2	…	Y1000,M
	第1000筆資料	X1000,1000,1	X1000,1000,2	…	X1000,1000,N	Y1000,1	Y1000,2	…	Y1000,M

圖 9　輸入特徵與輸出標籤向量示意圖

資料來源:本研究

3.4.2 模型訓練

本研究主要使用 matlab 建立模型並搭配不同的參數調整出最佳的模型，搭配 InceptionV3 預訓練之詳細模型架構如表 1 所示。圖 10 為神經網路架構圖，在訓練過程中使用 RMSE 作為 Loss 函數，將模型實際輸出與 Word2vec 期望輸出之間的均方誤差，並套用 RMSE 計算公式，以盡可能降低 RMSE 作為訓練目標。

表 1　InceptionV3 架構表

Layer	Patch size/units	stride	Input size
Conv	3*3	2	299*299*3
Conv	3*3	2	149*149*32
Conv padded	3*3	2	147*147*32
Pool	3*3	1	147*147*64
Conv	3*3	1	73*73*64
Conv	3*3	1	71*71*80
Conv	3*3	1	35*35*192
3*Inception	3*3	3	35*35*288
5*Inception	17*17	3	17*17*768
2*Inception	8*8	2	8*8*1280
Pool	8*8	2	8*8*2048
Global Average Pooling	logits	1*1*2048	
Dense	4000		
Dropout	0.2		
Dense	2500		
Dropout	0.2		
output	100		

資料來源:本研究

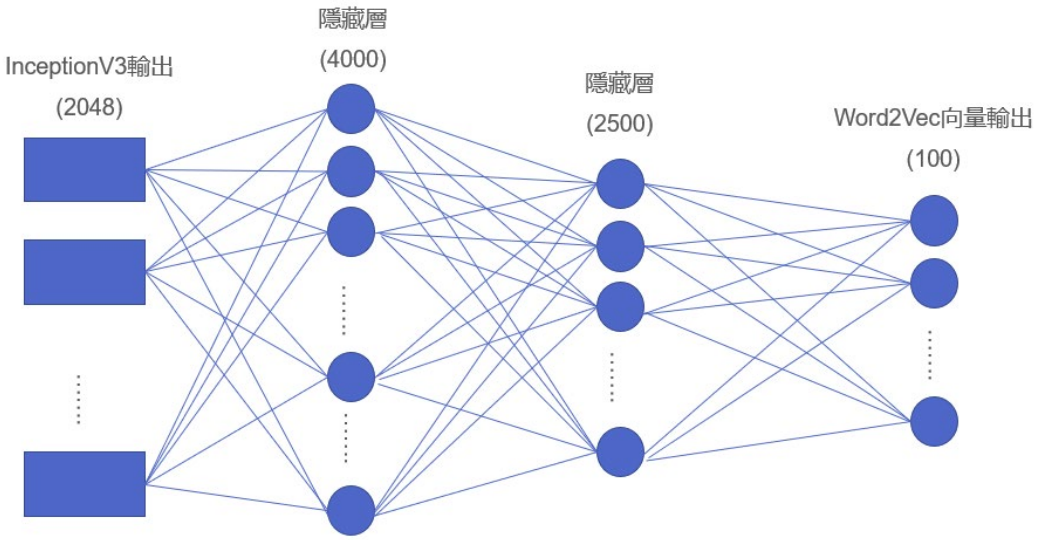


圖 10　神經網路架構圖
資料來源:本研究

3.5 圖像分類

訓練好 regression network 後會經過以下步驟將模型輸出之向量轉換成圖像類別:

1. 給任意一張圖片，將他交給 regression network 執行，令 regression network 輸出的結果為: [v1, v2, …, vM]

2. 接著，將[v1, v2, …, vM]與[Y1,1　Y1,2　…　Y1,M],…, [Y1000,1Y1000,2　…Y1000,M]比較相似度，其中，[Yi,1　Yi,2　…Yi,M]是第 i 個類別的詞嵌入向量

3. 如果[v1, v2, …, vM]與[Yk,1　Yk,2　…　Yk,M]的相似度最高(尤拉距離最小)，則將圖片辨識為類別 k，其中，[Yk,1　Yk,2　…　Yk,M]是第 k 個類別的詞嵌入向量

輸出向量與實際類別之示意圖如圖 11 所示，黃點代表模型輸出向量，紫點代表標籤詞遷入向量。以此圖為範例會將黃點歸類為紫點所代表的類別。

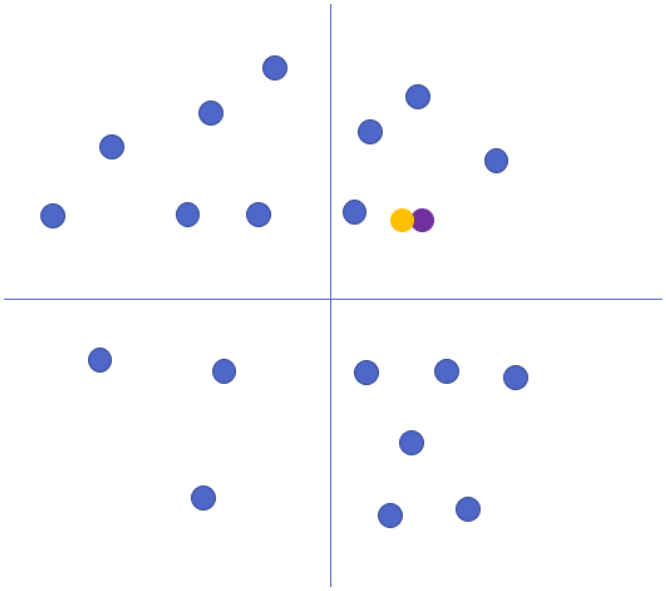


圖 11　輸出向量與實際類別向量比對示意圖
資料來源:本研究

3.5.1 靶心偏移

上小節提到，本研究之模型目的為最小化 RMSE，但是 RMSE 越小並不一定代表準確率越高。傳統的分類，期望結果是不可以變的，現在我們的方法，期望結果是可以調整的。透過微調訓練詞遷入向量，將預測正確的類別向量與模型的向量輸出互相靠近。並將預測錯誤的類別向量與模型輸出相互遠離，來達到透過靶心偏移將圖片的知識融入到文字的知識中，使得新的 Word2vec 同時包含文字與圖片的知識。具體做法與虛擬碼如圖 12 所示。

```
predicted_vec <- an array from YPredicted
word_vec <- an array from new_word2vec
for iter (1 to MaxNum) do
        for i (1 to Validation_data_num) do
                a <- YPredicted(i,ALL)
                b <- ones(1,1000) // 建立1000*100的矩陣，每一行都是第i各validation資料的預測結果
                predict_Y_result <- b * a
                dist <- predict_Y_result - word_vec  //計算預測結果與1000各類別的詞潛入向量的距離
                dist_r2 <- dist^2
                distance <- sum(dist_r2)
                [res,id] <- sort(distance); // 把預測結果與1000各類別的詞潛入向量的距離排序
                if Validation_data_label(i) == id(1) then
                        word_vec(1,ALL) <- word_vec(1,ALL) + lr*(YPredicted(i,ALL) - word_vec(1,ALL));
                        //若預測類別是正確，將原本的詞還入向量，往預測的方向移動，製作成新的詞還入向量
                else //top1_acc不正確
                        ind = find(where the correct answer in distance); //找出正確類別是排序距離後的第幾位，ind代表正確類別是排序第ind個
                        word_vec(ind,ALL)<- word_vec(ind,ALL) + wr*lr*(YPredicted(ind,ALL) - word_vec(ind,ALL));
                        //將正確答案的詞還入向量往預測方向移動，並乘以權重以增加移動幅度
                        for wrong_id (1 to ind-1) do
                        // ind代表正確類別是排序第ind個，所以ind之前的類別都是錯誤分類
                                word_vec(id(wrong_id),ALL) <- word_vec(id(wrong_id),ALL) + wr*lr*(i/ind)*(word_vec(id(wrong_id),ALL) - YPredicted(i,ALL))
                                // 將錯誤答案的詞還入向量往預測的反方向進行移動，並乘以權重以增加移動幅度
                        end for
                end if
        end for
end for
```

圖 12　靶心偏移虛擬碼
資料來源:本研究

靶心偏移與 MLP 網路訓練的具體步驟為以下兩點:

1.固定 Word2vec 向量，並視為 MLP 網路的期望輸出，來調整網路權重以降低 RMSE。

2.固定 MLP 網路權重，調整 Word2vec 向量(靶心偏移)以提高準確率。

透過以上兩個步驟交替執行，直到整體收斂為止。

肆、研究結果

本研究利用 ImageNet 中 ISLVRC_ 2017 年版資料集訓練機器學習並使用自然語言模型、轉移學習等方式來驗證使用詞嵌入向量作為類別的期望輸出向量，是否會使模型最終準確率提升。本章節將比較 1.使用不同層數、不同輸出空間的隱藏層之成效，2.比較不使用與使用靶心偏移演算法兩者之差距，3.比較目前有名之預訓練模型與本模型之差距。RMSE 為均方根誤差，作用為計算預估值與實際值之間的差異量，主要用於評估回歸型模型的準確率，故本研究使用 RMSE 作為模型評估指標之一，其計算公式如圖 13 所示。Top-1 與 Top-5 accuracy 是很常用於 Image Net 相關論文中的一種準確率指標，Top-1 代表的意義為模型預測機率最大者與正確答案一致才是為正確。Top-5 則是包含機率排序後之前五項類別，若其中包含正確答案即視為正確。為了能方便與近期熱門模型做相互比對，本研究採用 Top-1 accuracy、 Top-5 accuracy 評估不同模型方法之績效差異。

$$RMSE = \sqrt{\frac{SSE_W}{W}} = \sqrt{\frac{1}{W}\sum_{i=1}^{N} w_i u_i^2}$$

圖 13　RMSE 計算公式

資料來源: SAP Analytics Cloud 根均方誤差 (RMSE)

4.1 相異隱藏層之績效評估結果

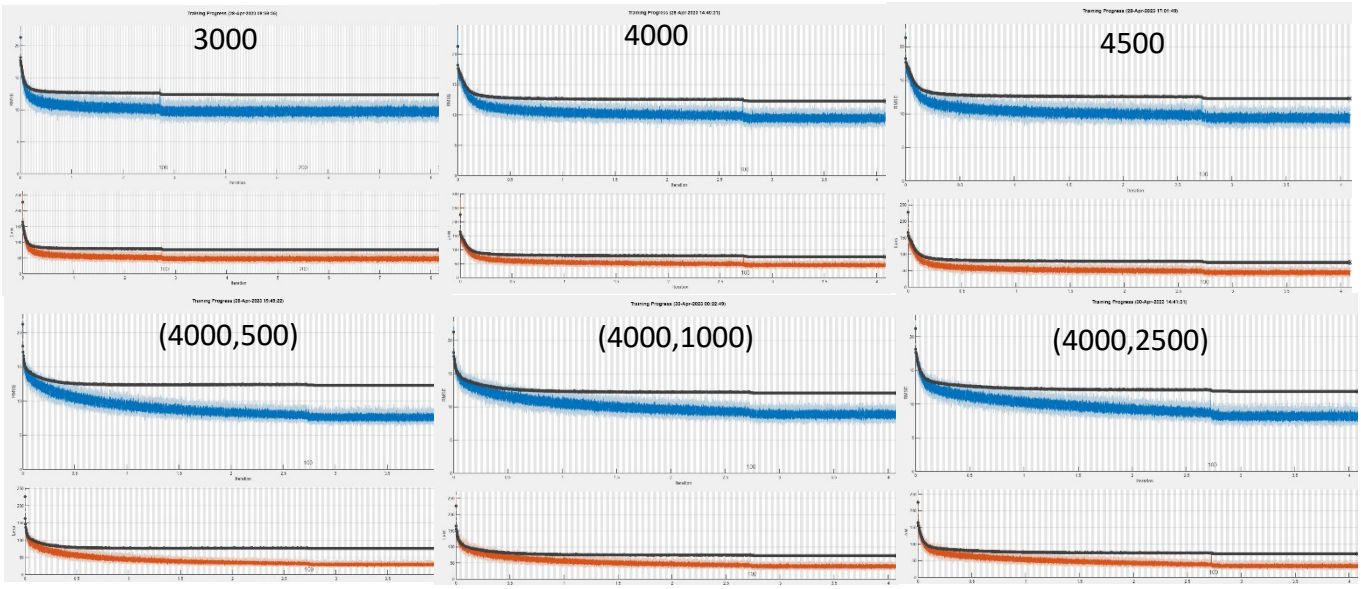
本研究除開隱藏層數與輸出空間、將其餘之參數依照以下規格設定: BatchSize :200、Epochs:150、activation: tanh、輸出向量:100。並比較一層:3000、4000、4500 與兩層(4000,500) 、(4000,1000) 、(4000,2500)之 MSE、Top-1 accuracy、 Top-5 accuracy 之差距。如表 2 所示、在隱藏層數為(4000,2500)時、RMSE 為 1.189422、Top-1 accuracy 為 0.625580、 Top-5 accuracy 為 0.681240 均高於其餘模型。故本研究將以兩層隱藏層(4000,2500)為基礎比較靶心偏移演算法之差距。表 3 為相異隱藏層之 Train、Test 曲線圖。上半部分為 RMSE，下半部分為 Loss，距離越遠代表 overfitting 越嚴重。

表 2　相異隱藏層之模型評估表

隱藏層數	RMSE	Top-1	Top-5
3000	1.237732	0.541020	0.603720
4000	1.228767	0.552520	0.614300
4500	1.225776	0.554760	0.616660
(4000,500)	1.235186	0.398900	0.419960
(4000,1000)	1.208456	0.479980	0.559880
(4000,2500)	1.189422	0.625580	0.681240

資料來源:本研究

表 3　相異隱藏層之 Train、Test 曲線圖



資料來源:本研究

4.2 靶心偏移演算法之類別分群結果

為了能夠確認執行靶心偏移演算法前與後的差別，本研究使用 K-mean 分群演算法將 1000 個圖片類別分為 20 群，並以透過各群類別數、圖形化以及使用 Calinski-Harabasz(CH)內部分群有效指標進行評估。一般分群有效性指標分為兩種: 一是外部標準，通過評估分群結果和參考標準的一致性來評價分群結果好壞； 另一種是內部指標，用於評價同一種分群演算法在不同分群數條件下分群結果的陳度，通常用來確定數據集的最佳分群數。而本研究使用的 CH 分群指標是基於數據統計訊息的指標，會根據數據集本身和分群結果的統計特徵對分群結果進行評估，並根據分群結果的優劣選取最佳群數。CH 指標的數字越大代表類別自身越緊密，類別與類別之間月分散，分群效果越好。[A Dendrite Method for Cluster Analysis]。首先每群的類別數如表 4 所示，附有網底的代表該群的類別數過多或過少，可以發現執行靶心偏移前的分群有 3 群的類別數量僅為 1，且第 5 群的類別數量超過一半，這些狀況執行靶心偏移後狀況有明顯的改善。

表 4　靶心偏移演算法之分群類別數

編號	1	2	3	4	5	6	7	8	9	10
前	1	64	35	23	513	39	17	29	41	1
後	26	5	25	14	34	33	48	20	24	25
編號	11	12	13	14	15	16	17	18	19	20
前	1	37	28	21	29	11	23	75	10	2
後	36	434	23	20	40	35	26	58	24	50

資料來源:本研究

關於 CH 內部分群指標評估如表 5 所示，可以發現在分群前的 CH 指標最高為 5 群的 40.32，從數字的意義上表示使用靶心偏移前的各類別之間過於緊密，很難看出類別間的差異性。而在使用靶心偏移後整體的分數有往上提升，且最高的 5 群從 40.32 上升至 152.02，代表靶心偏移能夠為類別的分類帶來正向的影響。

表 5　靶心偏移演算法之 CH 分群指標評估

群數	5	10	15	20	25
前	40.32428	25.75226	21.68061	19.41594	17.38697
後	152.0246	79.81424	57.65276	45.54109	37.88257

資料來源:本研究

關於圖形化之分群結果如圖 14 與圖 15 所示。圖 14 為執行靶心偏移前的圖像化分群，可以發現各類別非常緊密且集中於中心，絕大部分被分類為某一群難以辨別。圖 15 為執行後的分群結果，可以看到由原本密集的類別被劃分為兩大群，且兩大群中皆有不同的小群體組成，由上述的論點證明靶心偏移能為類別分類帶來正向的影響。

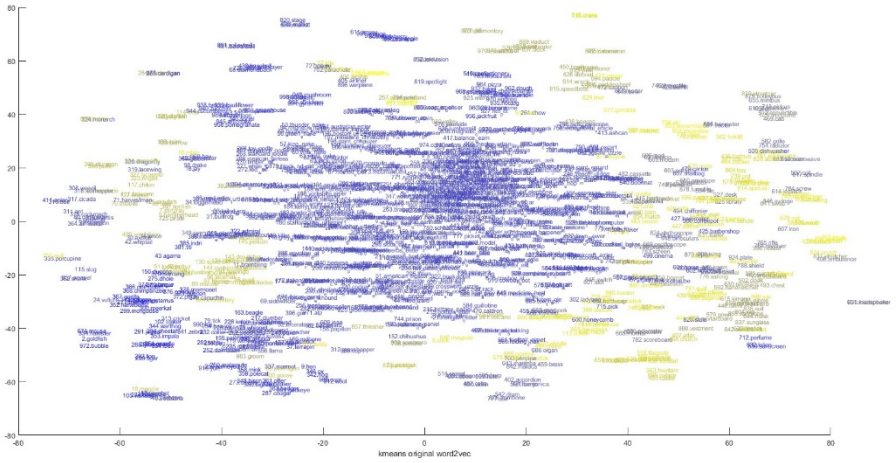


圖 14　靶心偏移前之分群結果圖
資料來源:本研究

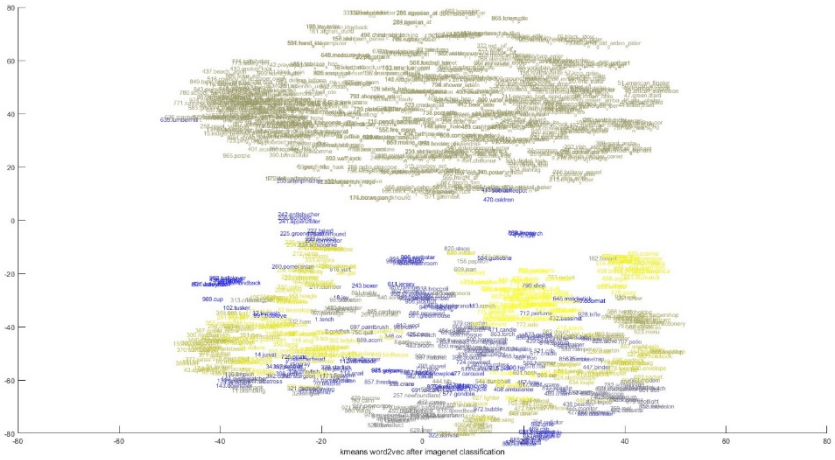


圖 15　靶心偏移後之分群結果圖
資料來源:本研究

4.2.1 靶心偏移演算法之績效評估結果

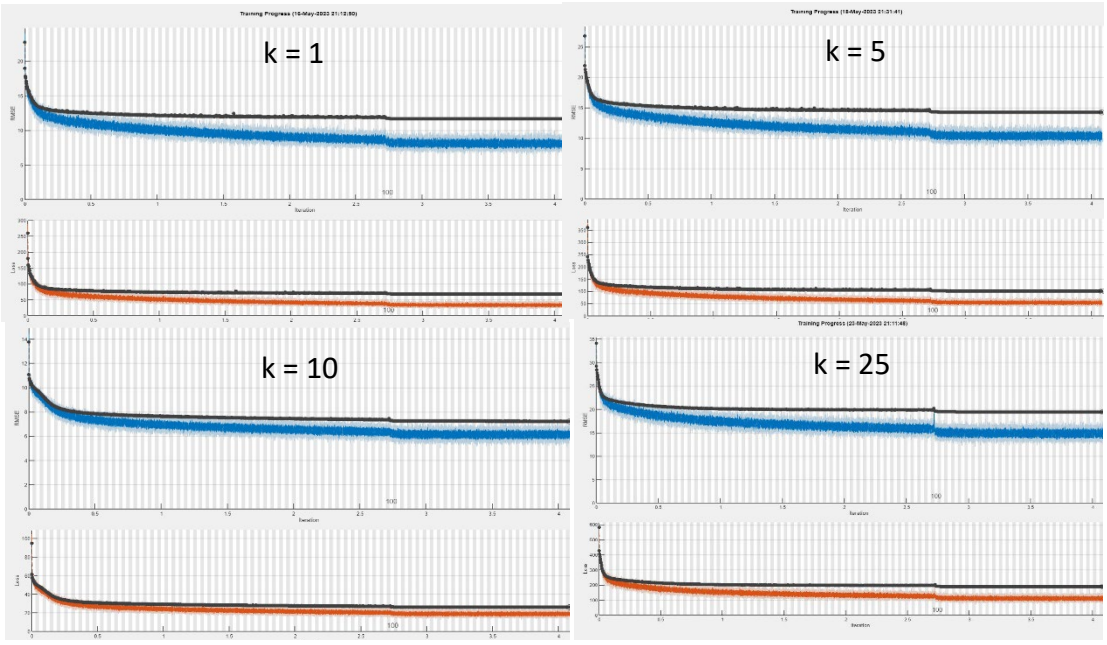
本研究將固定以下參數: BatchSize :200、Epochs:150、activation: tanh、輸出向量:100、隱藏層數:(4000,2500)。並比較有無使用靶心偏移演算法之差距。如同表 6 所示，K 作為訓練與靶心偏移交替次數。在 K = 25 下 RMSE 為 1.947281、Top-1 accuracy 為 0.781440、 Top-5 accuracy 為 0.871140 均高於其餘模型。表 7 為靶心偏移演算法之 Train、Test 曲線圖。上半部分為 RMSE，下半部分為 Loss，距離越遠代表 overfitting 越嚴重。

表 6　靶心偏移演算法之模型評估表

靶心偏移	RMSE	Top-1	Top-5
無	1.189422	0.625580	0.681240
K = 1	1.168064	0.756920	0.848660
K = 5	1.428162	0.772760	0.865120
K = 10	0.724436	0.771980	0.874680
K = 25	1.947281	0.781440	0.871140

資料來源:本研究

表 7　靶心偏移演算法之 Train、Test 曲線圖



資料來源:本研究

4.3 熱門預訓練模型之績效比較

本研究將比較 AlexNet、GoogLeNet、ResNet-18、VGG-16、Inception v3 與本研究模型，並以 Top-1 accuracy 、 Top-5 accuracy 作為評估指標。如表 8 所示，在 Top-1 accuracy 方面本研究之模型高過 AlexNet、GoogLeNet、ResNet-18、VGG-16 但略低於 Inception v3，Top-5 accuracy 方面 則除了 AlexNet 其餘皆高於本研究之成果。

表 8　熱門預訓練模型之績效比較表

模型名稱	Top-1	Top-5
AlexNet	63.3%	84.6%
GoogLeNet	68.3%	88.7%
ResNet-18	72.05%	91.75%
VGG-16	74.4%	91.9%
Inception v3	78.8%	94.4%
本研究	78.14%	87.11%

資料來源:本研究

伍、結論

本研究於第一章、第二章說明研究動機、研究目的、相關文獻與背景知識，於第三章說明研究方法與研究架構、資料來源、模型架構，在第四章對本研究的實驗結果針對各模型的最終效能進行比較，以及針對靶心偏移演算法進行說明與證明效用，本研究將於本章節對上述章節進行統整、研究限制以及研究建議對未來學者提供相關參考。

5.1 結論

本研究使用 ISLVRC 圖片資料集與 wiki-en 語料庫進行自然語言學習、轉移學習等模型訓練，以證明使用類別之詞遷入向量是否會幫助圖片進行學習。實驗資料包含 1000 個不同種類之物品、動物圖片與超過兩千萬篇以上之文章提供模型學習。本研究使用兩種模型:Word2vec 語言模型與 InceptionV3 遷移學習並將兩者合二為一，並透過靶心偏移演算法針對原模型詞遷入向量進行優化，並反覆訓練直至模型收斂。模型預測結果為類別對應之詞遷入向量，於訓練完成後進行轉換與分析，最終對所有模型績效進行比較。

在使用靶心偏移演算法前之測試結果中，績效最佳的為使用隱藏層(4000,2500)，其 RMSE 為 1.189422、Top-1 accuracy 為 0.625580、 Top-5 accuracy 為 0.681240，均高於其餘模型。在使用靶心偏移演算法後之測試結果中，績效最佳的為權重 K=25，其 RMSE 為 1.947281、Top-1 accuracy 為 0.781440、 Top-5 accuracy 為 0.871140，與未採用靶心偏移演算法前 Top-1 accuracy 相差了 0.15586。綜合上述實驗，可以得知使用 Word2vec 詞遷入向量作為模型訓練類別並搭配靶心偏移演算法能夠藉由針對詞遷入向量的微調來獲得模型準確率之提升。

5.1.1 相關文獻結果比對

Wonjae Kim 等學者創造出使用 transformer 處理視覺特徵並在預訓練過程使用 whole word masking 和圖像增強技術在一般圖像學習任務上之 R@1(Top-1)最佳為 64.4，Zhicheng Huang 等學者位的解決 ROI 只能獲取到一部分特徵的缺陷作者使用了一套 VD-base embedding(Visual Dictionary)模組來針對圖片進行處理，其研究結果在圖文偵測上 R@1(Top-1)最佳為測試資料量 1K 之模型，準確率為 73.5。與過往研究相比本研究能獲得更高的準確率。因此，本研究足以說明使用類別之詞遷入向量並配合靶心偏移能夠帶來更好的績效。

5.2 研究限制

本研究將整體之研究限制統整為以下兩點:

1. 為了加快訓練時間與減少硬體負擔，本研究所使用之圖像訓練資料僅涵蓋已標注資料，與本研究模型架構之來源 InceptionV3 上所使用之樣本與訓練時間有明顯之差距，因此略低於 InceptionV3 之模型績效。
2. 本研究所使用之語料庫僅透過龐大的文字量作為基底，並沒有針對類別物件進行語詞上的特化。從第四章之分群結果可以發現，儘管已經套用靶心偏移，仍然有一群所涵蓋的類別將近於總數之一半。

5.3 建議

由於本研究與現今熱門預訓練模型相比，所耗費之時間與擁有之計算能力皆不足，若能提高運算能力亦或是使用全部的圖片資料作為訓練資料，便能提高整體模型績效。在語料庫的使用上可以搭配 YouTube 圖片或 Google 文章關鍵字搜尋來獲取類別相關文章達到針對類別做特化，以提升模型準確度與研究可信度。

參考文獻

1. Wonjae Kim et al, 2021, “ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision” , PMLR , vol. 139,, ,Pages 5583-5594 , July

2. Zhicheng Huang et al, 2021, “Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning” , arXiv preprint arXiv: 2104.03135, Apr 07

3. Abhisek Kundu et al, 2019, “K-TanH: Efficient TanH For Deep Learning” , arXiv preprint arXiv:1909.07729 , Sep 17.

4. Tomas Mikolov et al, 2013, “Efficient Estimation of Word Representations in Vector Space” ,arXiv preprint arXiv:1301.3781 , Jan 16.

5. Christian Szegedy et al, 2014, “Going Deeper with Convolutions” ,2015 IEEE Conference on Computer Vision and Pattern Recognition. Pages 1-9 Jun 12

6. Christian Szegedy et al, 2016, “Rethinking the Inception Architecture for Computer Vision” , 2016 Proceedings of the IEEE conference on computer vision and pattern recognition Pages 2818-2826 Jun 27.

7. Nitish Srivastava et al.,2014,“Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, Journal of Machine Learning Research Vol.15,Pages 1929-1958, Jun 14

8. T. Caliński et al 1974 “A dendrite method for cluster analysis”, Communications in Statistics, Vol.3, Pages 1-27, Jun 01

9. Rumelhart, D. et al,1986. “Learning representations by back-propagating errors” Nature Vol 323, Pages 533–536, Oct 09

10. Nair, Vinod et al. 2010. “Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair”. Proceedings of ICML. Vol.27. Pages 807-814, Jun 21

11.Yu-Hsien Yeh.,2018,“Machine Learning — Transfer Learning (遷移學習)” , medium from:https://medium.com/@yuhsienyeh/machine-learning-transfer-learning-%E9%81%B7%E7%A7%BB%E5%AD%B8%E7%BF%92-5095f8a14367 , Jul 18

12. Garson, et al., 2018 ,“"Connectionism", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition) ” from: https://plato.stanford.edu/archives/fall2018/entries/connectionism/ , Sep 21