在圖像分類任務中設計深度學習模型架構與網路超參數應用於 Few shot learning

I. 研究動機及目的

構建一個僅使用少量已標註資料進行訓練，可以快速應用到未見過的新任務中又表現良好的模型是元學習(meta learning) 的一個挑戰。在過去的研究中，透過設計一個孿生網路架構，並基於幾個常用的損失函數之下嘗試提議用於孿生網路上更有效的損失函數，實驗結果顯示，使用調適後的基本網路和對比損失函數（Contrastive Loss）、Adam 優化器和 ReLU 激活函數，在 MNIST、LFW、AT&T 資料集上，以小資料量（基於 One-Shot 學習架構）和較少的網路參數量，分別在測試資料上，達到 97.27%、76%與 75.63%的識別準確率，且訓練收斂速度較快。

在本計畫的研究範疇裡，將延續過去之研究主軸圖像分類模型架構之設計，實驗提升辨識準確率的方法。相關的研究包含，[1-2]導入注意力模塊，提升模型特徵提取效能，[3]研究模型架構與損失函數設計、[4-5]將孿生網路的相似度概念以交叉注意力的方式呈現、[6-12]研究導入語義特徵輔助圖像模型辨識。

關鍵詞：Few shot Learning, MutiModel learning, natural language supervision, vision Language Model, Cross Attention

II. 研究方法

2-1 卷積注意力模塊 (Convolutional Block Attention Module, CBAM)

2-1-1 相關研究 [1-2]

卷積注意力模塊由通道注意力模塊與空間注意力模塊組成，分別應用注意力於通道與空間兩個維度中。通道(channel)注意力模塊：在特徵圖的每一個通道中，皆有注意力檢測器，實施機制為在單一空間中進行平均池化與最大池化，得到兩個 $1 \times 1 \times C$ 的通道描述，再送入兩層的 MLP，得到的特徵相加後經過 Sigmoid 得到注意力的權重係數；空間(Spatial)注意力模塊：與通道注意力模塊相反，分別進行單通道的平均池化與最大池化，得到兩個 $H \times W \times 1$ 的空間描述，後續處理程序皆相同。

CBAM 可以靈活的插入 CNN 的 GAP 層後，提高重要部分特徵的權重，並抑制不必要的特徵，且可與 CNN 一起進行訓練，在不同的分類資料集中，加入 CBAM 模型表現皆有一定的提升。

2-1-2 研究方法

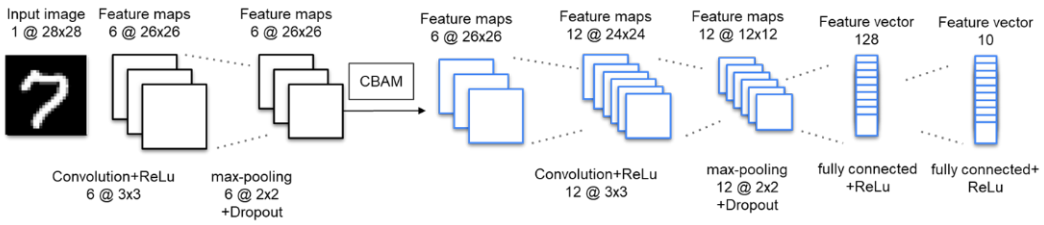

圖 1 加入 CBAM 後的模型架構，在模型第一個池化層後加入 CBAM，對特徵進行注意力機制選取。

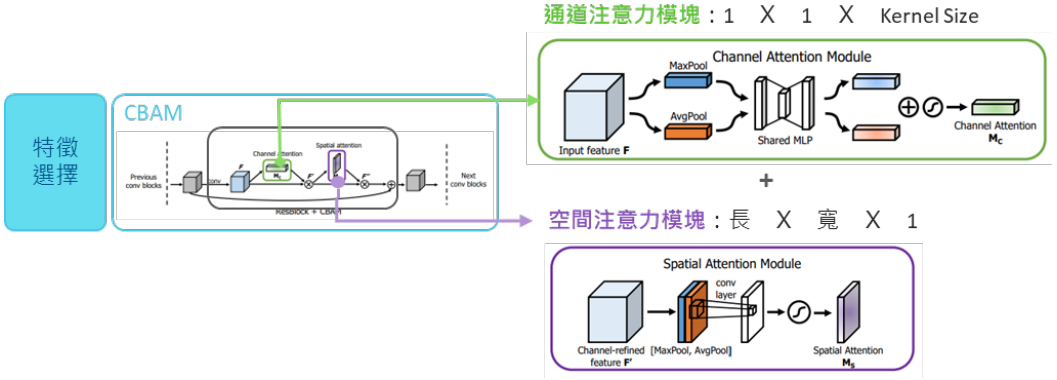

圖 2 CBAM 之組成，參考[1]

2-1-3 問題討論

實驗結果模型表現使用 contrastive loss 或 focal loss 皆沒有上升，且訓練時準確率隨著 epoch 增長而下降，注意力模塊或模型架構可能存在問題。

2-2 自適應損失函數 (Adaptive Loss Function)

2-2-1 相關研究[3]

weighted contrastive loss：具有較大距離和較小距離的負樣品的正樣品被認為是孿生網絡的難分類樣品(hard samples)，在訓練中更重要。但是，傳統孿生網絡中的樣本對標籤穩定。為了增加訓練中硬樣本的比例，我們設計了對比損失的權重，以專注於這些難分類樣品，函數表示如下：

$$L_1(y, d_{pos}, d_{neg}) = y \times (1-w) \times d_{pos} + (1-y) \times w \times \max(0, m - d_{neg}) \quad w = \frac{L_1 \cdot L_2}{||L_1||\,||L_2||}$$

L_1、L_2：為從兩個特徵圖上獲取的中心向量；w使用餘弦距離度量數入對相似度；d_{pos}為正樣本輸入對的特徵距離；d_{neg}為負樣本輸入對的特徵距離。

adaptive cross-entropy loss：為了平衡兩平行輸入特徵，使用自適應參數（α 和 β）融合兩個分別學習自兩分支中的特徵向量，則 adaptive cross-entropy loss 的函數表示如下：

$$L_2(\hat{y}_i, y_{iP_1+iP_2}) = -\sum_{i=0}^{N} y_{iP_1+iP_2} \log(\hat{y}_i) \quad y_{iP_1+iP_2} = Softmax(\alpha \times y_{iP_1} + \beta \times y_{iP_2})$$

y_{iP_1}、y_{iP_2}為兩輸入樣本的特徵向量；α 和 β 初始值設為 0.5。

2-2-2 研究方法

使用 weighted contrastive loss (L1) 訓練兩輸入之間的相似度辨識能力，而 adaptive cross-entropy loss (L2) 學習分類能力。$Loss = L_1(\textit{相似度損失函數}) + \lambda L_2(\textit{分類損失函數})$

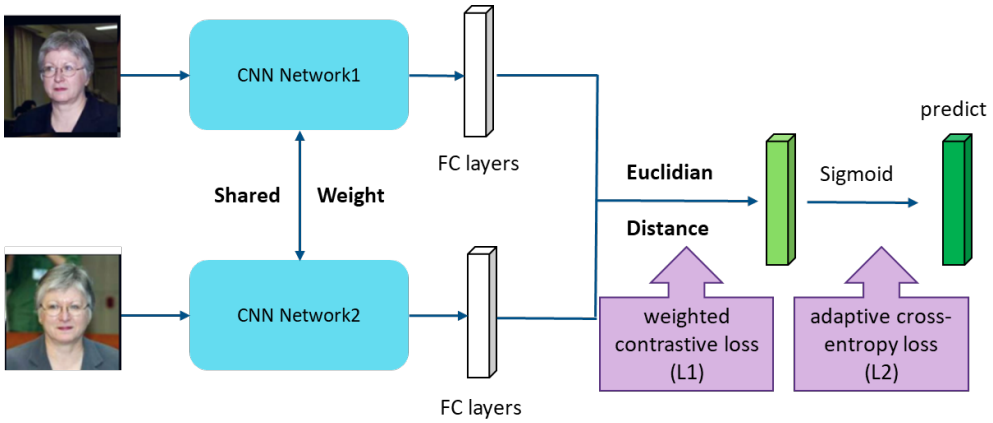

圖 3 孿生網路與雙損失函數的架構

2-3 Cross Attention [4-5]

2-3-1 相關研究

過去研究之孿生網路架構，是基於特徵間相似度並使用歐式距離作為度量，過曲使用過注意力機制用於網路中進行特徵選擇，而交叉注意力則能取代孿生網路架構，將 support set 與 query image 的特徵 concat 起來，經過交叉注意力，將 query image 映射到各個不同類別的 support set 的特徵空間中，得到基於不同類別的 query image 注意力；將 support set 裡的圖像映射到 query image 的特徵空間中，得到基於 query image 的 support set 注意力。再將 query image 的特徵圖分別與各個不同類別的 support set 相乘，取相似度最大的類別。

損失函數的計算則進行兩種任務， Nearest Neighbor classification 利用 query image 與各不同類別 support set 的圖像相似度進行聚類，判斷最相似類別； Global class classification 學習網路經過分類器(全連接層)的預測類別，判斷是否分類正確。

2-3-2 研究方法

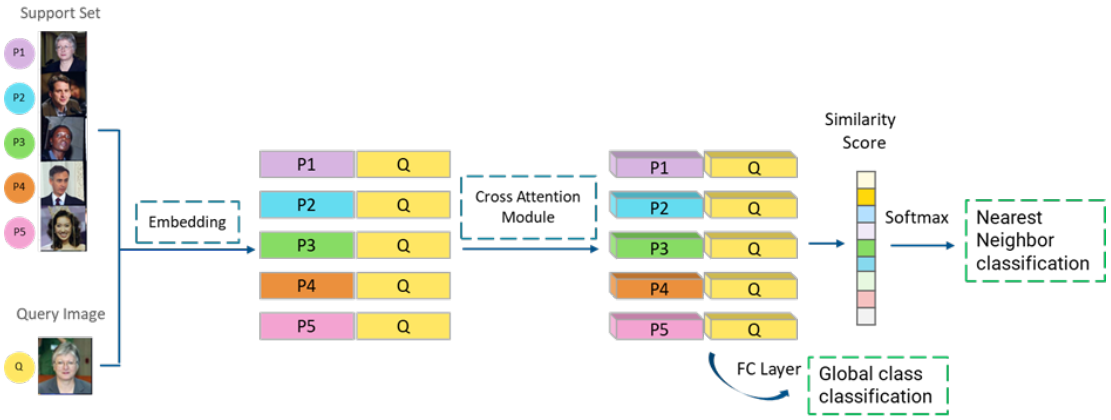

圖 4 基於 cross attention 的網路架構，參考[4]

2-3-3 問題討論

使用作者提供的程式，撰寫自定義資料集(LFW)的 loader 後，模型表現得到很好的提升(83%)，且訓練速度加快很多(60 個 epoch)，而訓練所花 GPU 記憶體容量並沒有增加(12G)。

圖 5 使用圖 4 架構訓練得到的預測結果。Label 表示該圖片的類別代碼，support set 中標示為 1 的圖片與 Test set 中預測為 1 的圖片確實為同一類。

2-4 基於多模態學習建構之視覺語言模型 (Vision Language Model)

在影像分類任務中，常見的訓練方式是使用影像特徵作為模型輸入，但在實際訓練中，完全靠影像特徵來訓練，模型泛化性差且弱於分辨差異很小的種類。導入多模態學習(MutiModel Learning)，採用影像特徵結合語義特徵，建構視覺語言模型(vision Language Model)，在提升模型表現的同時，增加模型泛化性。

2-4-1 相關研究[6-13]

對於視覺語言模型，主要包含三個模塊，圖像編碼器、文本編碼器以及合併兩者的學習技巧。文本編碼器一般使用可以帶入預訓練模型的 Tramsformer，因其在文本任務中表現良好，且不用另外耗費大量時間對其進行訓練；圖像編碼器則可以使用與 Transformer 架構較相似的 Vision Transformer 或是 CNN 架構的 NFNet，根據不同的學習策略有不同的對圖像編碼的模式。在圖像及文本編碼器上，皆使用預訓練模型對不同訓練任務作微調，這些預訓練模型擁有良好的遷移性能，可以適用於不同的訓練目標。合併兩者的學習策略是藉由幾種關鍵元素進行設計，損失函數以及模型框架，根據不同的設計方法，目前的學習方法有以下幾種：

- 單一流(stream)：視覺特徵嵌入語言模型中，視覺模型與語言模型呈單線序列訓練。
  - 前綴語言模型(PrefixLM)[7]：將圖像 token 作為文本 token 的前綴來學習視覺和語言的聯合訓練。
  - 基於交叉注意力的多模態融合(Multi-modal Fusing with Cross Attention)[8]：將視覺注意力的特徵使用交叉注意力融合到語言模型中，而不是直接作為前綴更改語言模型。
- 雙流(two stream)：視覺與語言模型同時萃取特徵，呈平行訓練。
  - 對比學習(Contrastive Learning)[6]：將圖像與文本的特徵在聯合空間中比對，使用個文本向量取代原全連接層，將圖像特徵在文本向量中做對比。
  - 遮罩語言模型 (Masked-Language Modeling，MLM) / 圖像文本匹配 (Image-Text Matching，ITM)[9]：使用遮罩語言模型預測遮罩圖片對應的語意分類上的分布；圖像文本匹配預測圖像文本是否對齊。
  - 無訓練[10]：相似的圖像也會有相似的文本特徵，用訓練好的視覺單模

態模型創建一個相對特徵表示空間，在該空間進行文本相似性的搜索。

2-4-2 研究方法

提出一種使用多模態模型，參考對比[11]與雙流[12]學習模式，研究不同模態間的特徵融合，讓模型得以執行圖像分類與標框任務。模型由視覺與語言兩個流組成，然後使用視覺特徵表示模塊將兩者特徵融合。

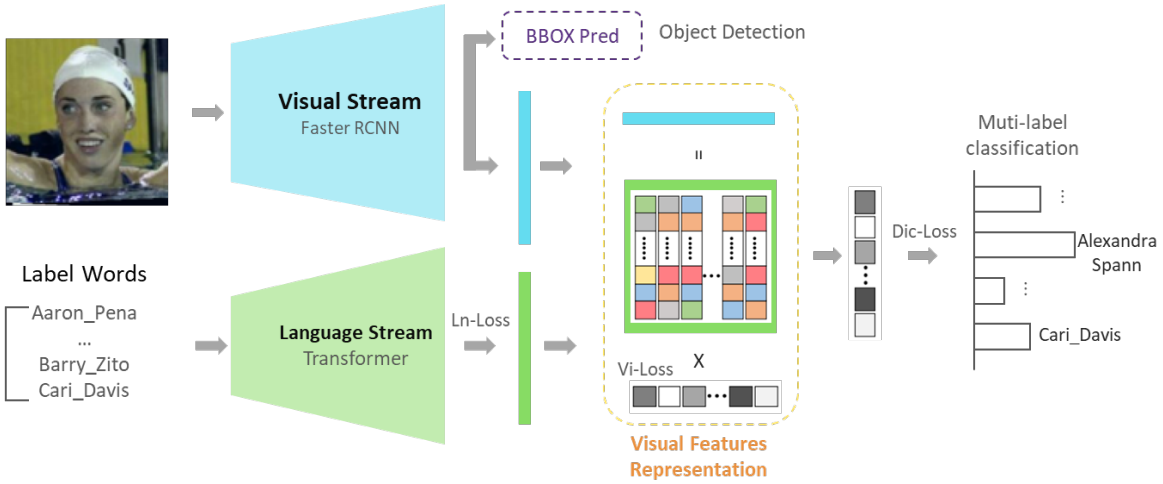

圖 6 視覺語言模型之架構，參考[11]

- 視覺流（Visual Stream）：由 Faster R-CNN[13]獲取圖像多個目標區域的特徵與標框預測結果。
- 語言流(Language Stream)：使用預訓練 Transformer 模型，編碼器將文本投映到視覺空間，解碼器再將文本特徵投映回語言空間。使用餘弦距離來度量文本編碼解碼前後相似程度。
- 視覺特徵表示模塊（Visual Feature Representation）：使用學習到的模型，將文本特徵來表示覺流獲取到的圖像特徵。
- 損失函數（Loss Function）：

$$\frac{(\text{視覺特徵重構損失}(\text{Vi}-\text{Loss}，\text{歐式距離})+\text{文本分類損失}(\text{Dic}-\text{Loss}，\text{cross entropy}))}{(\text{語言特徵重構損失}(\text{Ln Loss}，\text{餘弦距離}))}$$

III. 參考文獻

[1] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[2] Takimoto, Hironori, et al. "Anomaly detection using siamese network with attention mechanism for few-shot learning." Applied Artificial Intelligence 36.1 (2022): 2094885.

[3] Xue, Zhaohui, Yiyang Zhou, and Peijun Du. "S3Net: Spectral–spatial Siamese network for few-shot hyperspectral image classification." IEEE Transactions on Geoscience and Remote Sensing 60 (2022): 1-19.

[4] Hou, Ruibing, et al. "Cross attention network for few-shot classification." Advances in neural information processing systems 32 (2019).

[5] Shao, Huikai, et al. "Few-shot learning for palmprint recognition via meta-Siamese network." IEEE transactions on instrumentation and measurement 70 (2021): 1-12.

[6] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
[7] Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.
[8] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in Neural Information Processing Systems 35 (2022): 23716-23736.
[9] Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in neural information processing systems 32 (2019).
[10] Norelli, Antonio, et al. "Asif: Coupled data turns unimodal models to multimodal without training." arXiv preprint arXiv:2210.01738 (2022).
[11] Zhou, Fengtao, Sheng Huang, and Yun Xing. "Deep semantic dictionary learning for multi-label image classification." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 4. 2021.
[12] Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." arXiv preprint arXiv:1908.07490 (2019).
[13] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).