CVPDL Homework 2

Generic Object Detection

1. (10%) Current transformer-based object detection models are primarily developed based on DETR. However, we all know that DETR[1] has a drawback, which is slow convergence speed. Please describe why DETR series models have slow convergence speed.

首先，造成 DETR 收斂速度慢的原因可以先分為幾個原因，初始化問題、二分匹配策略造成的匹配不穩定，大規模的 transformer 架構等，而最主要的原因似乎是在於初始化的問題，因為在 decoder 中，並不像 encoder 一樣有位置編碼輸入，而是可以學習的嵌入向量(learnable queries)，在剛開始計算 cross attention 時，大多數的 decoder embeddings 中的向量都會被投射到空間中相同位置，所以因為沒有位置先驗，DETR 模型就需要花費較長的時間來學習生成更好的 attention map。

2. (10%) Two well-known papers, DAB-DETR[2] and DN-DETR[3], approach the issue of slow convergence in DETR from different viewpoints. Please describe how they address the issue.

DAB-DETR: DETR 的 attention map 中，一個 query 要關注多個中心區域，於是乎 DAB-DETR 就加入動態的 Anchor Box，提供位置先驗，並且加入物體的尺度訊息，寬和高，而且在 decoder 的地方，位置先驗是 learnable queries，所以 Anchor Box 理論上是可學習的，且又加上 MLP 實現 coarse-to-fine，因此，DAB-DETR 透過引入動態 Anchor Box，提供了一種有效的解決 DETR 慢收斂問題的方法。

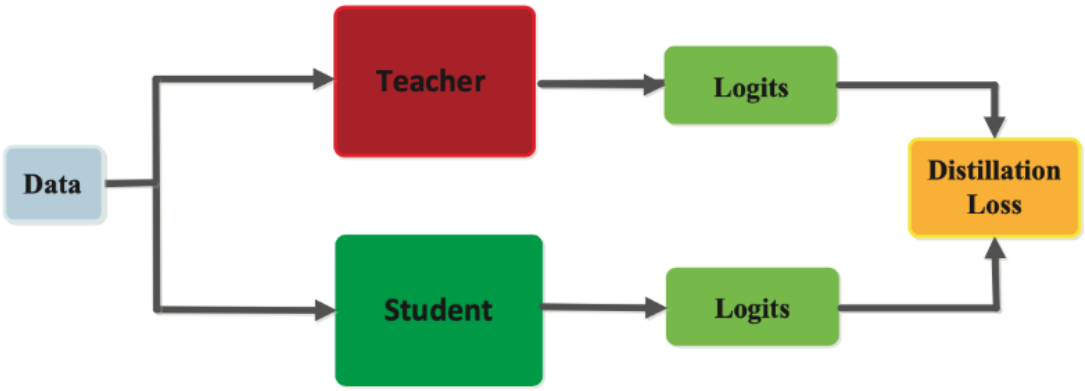
DN-DETR: DN-DETR 主要是通過引入 DeNoising training 來加速 DETR 的收斂，具體來說，DN-DETR 的 DeNoising training 是這樣運作的，首先，對 ground-truth 的邊界框（bounding boxes）添加一些噪聲，然後將這些帶有噪聲的邊界框輸入到 Transformer 解碼器中，訓練模型去重建原始的邊界框。而這種方法的好處是，由於添加的噪聲都很小，所以模型比較容易根據這些噪聲輸入去預測對應的 ground-truth，從而降低了學習的難度。不過有個地方要注意的是 DN-DETR 的 DeNoising training 只在訓練時需要，而在推理時是去掉的，並不會給最終模型的實際應用增加負擔。同時，這個任務的引入也不會改變模型結構，能夠與 DETR 系列的模型很好地兼容。因此，DN-DETR 能夠有效地改善 DETR 的收斂問題。

Practical Issue 1: Knowledge Distillation

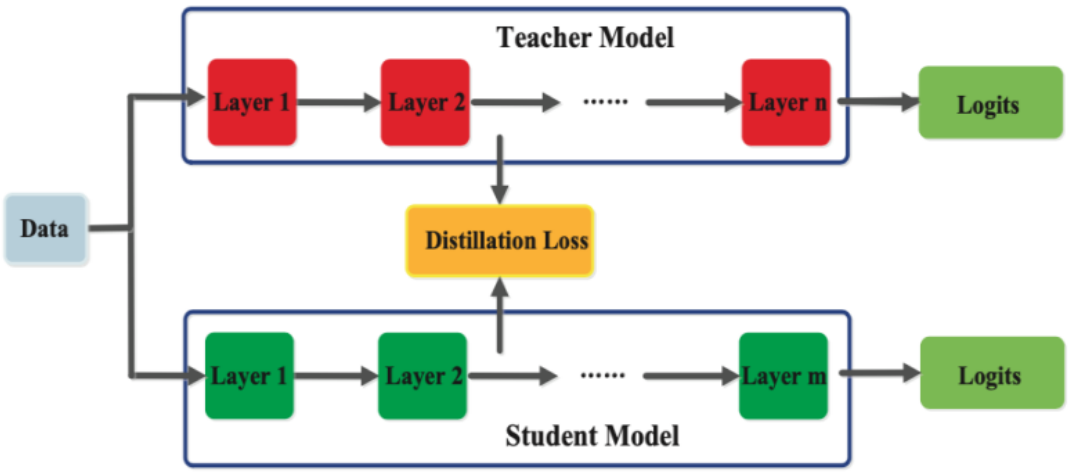
1. (10%) Briefly describe the main differences and advantages between distilled from logits and distilled from intermediate layers.

Knowledge distillation 是利用像是老師教學生的概念，將大的 Teacher model 作為小的 Student model 模仿學習的對象，並利用老師的 Knowledge 來獲得相似或更高的準確性

distilled from logits: 這種方法是使用 Teacher model 的輸出層（logits）作為 Student model 的目標，最小化 logit 之間的差異。這種方法的優勢在於，由於 Distillation loss 在訓練過程中被最小化，學生模型將能夠更好地做出與老師相同的預測。



distilled from intermediate layers:這種方法是利用 Teacher model 的中間層作為 Student model 學習的目標，這種方法的優勢在於，它可以讓 Student model 學習到 Teacher model 的中間表示，獲得更豐富的 Knowledge。



2. (15%) How to perform knowledge distillation from logits? Please find a paper from the given top conference list shown in Lecture#1 from 2021 or later discussing certain approaches performing distillation from logits and briefly describing its methodology.

1.需要一個已經訓練好的 Teacher model。

2.創建一個新的 Student model。這個模型通常比 Teacher model 小，而且未經訓練。

3.衡量 Student model 的 logits 與 Teacher model 的 logits 之間的 loss。

4.使用包含這兩個之間的 loss 來訓練 Student model。

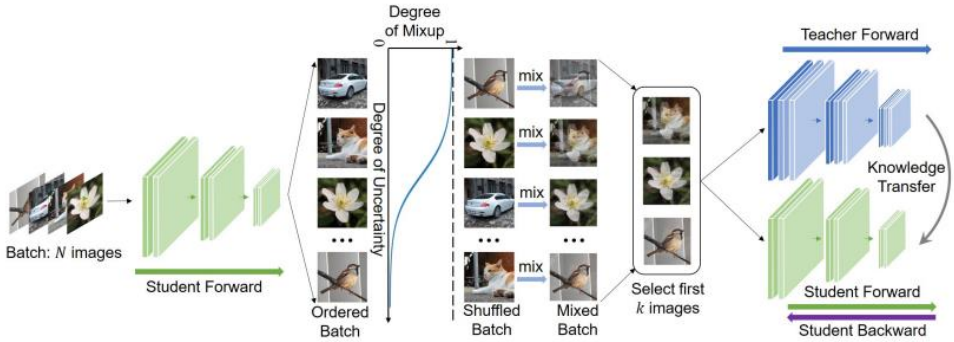
Paper: Computation-Efficient Knowledge Distillation via Uncertainty-Aware Mixup

1. Uncertainty Sampling Strategy：我們會將一個批次中的所有圖像輸入到 Student model，並獲得他們的不確定性，不確定性抽樣策略用於評估每個訓練樣本的訊息量。具體來說，一個學生不能很有效的分類的樣本是一個複雜的樣本，它可以在查詢時為模型帶來更多的資訊。這種策略有助於確定哪些樣本最有可能從教師網絡中獲取有用的資訊，並被選擇進行 Mixup。

2. Adaptive Mixup：這個方法將自適應混合應用於不確定的樣本以壓縮知識，將兩個圖像的內容壓縮到一個圖像中。這篇 paper 希望一個混合的圖像可以傳遞比一個正常圖像更多的 knowledge。然而，壓縮通常會帶來資訊的損失。以像素為單位的合併導致兩個圖像之間的互相混淆。每個圖像中的特徵都被混合圖像破壞，使得合成的圖像模糊並且在語義上無意義，而為了利用壓縮效果，同時減少對資訊樣本的破壞，以自適應的方式應用 Mixup 操作

在這篇 paper 中主要就是透過減少冗餘的概念優化 knowledge distillation 過程中的計算成本。傳統的 knowledge distillation 方法中存在的冗餘主要在於過度學習簡單的樣本。這種過度學習會導致計算資源的浪費，因為對於那些 Student model 已經能夠很好處理的簡單樣本，再次從 Teacher model 中獲取相同的 Knowledge 並沒有太大的必要。

因此，總結來看這個方法通過結合不確定性和混合，減少了冗餘，並更好的利用了對 Teacher model 的每次查詢。簡單來說，首先就是要將所有的訓練樣本輸入到學生網絡，並獲得他們的不確定性。然後，根據不確定性對樣本進行排序，並選擇最具資訊量的樣本進行混合操作。這樣，每次查詢 Teacher model 時，都能夠從兩個樣本中獲取更多的知識，提高了計算效率。



3. (15%) How to perform knowledge distillation from intermediate layers? Please find a paper from the given top conference list shown in Lecture#1 from 2021 or later discussing certain approaches performing distillation from intermediate layers and briefly describing its methodology

1.選擇 Teacher model 和 Student model

2.選擇中間層

3.進行 Knowledge Distillation：將教師模型的中間層的輸出作為目標，訓練學生模型的中間層來接近這些輸出。

4.評估 Student model：在訓練結束後，評估 Student model 的性能。如果學生模型的性能達到了一定的程度，則 Knowledge Distillation 就完成了。

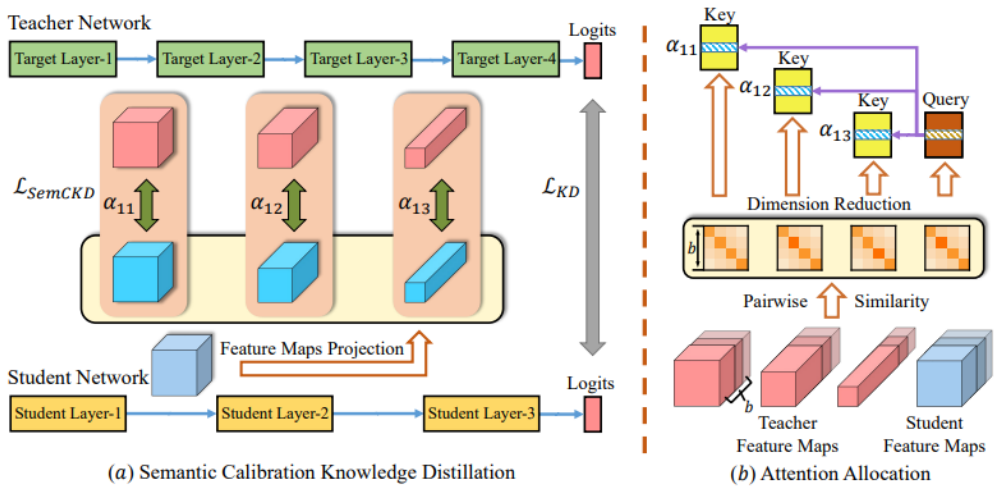
Paper: Cross-Layer Distillation with Semantic Calibration

這篇 paper 主要解決了在 Knowledge Distillation 過程中，Teacher model 和 Student model 之間的層次語義不匹配問題，他們提出了一種名為"Semantic Calibration for cross-layer Knowledge Distillation"（SemCKD）的方法，這個方法可以自動為每一個 Student model 的層分配適當的 Teacher model 的目標層。主要是通過注意力機制來實現的，在這種方法中，每一個 Student model 的層都會從多個 Teacher model 的層中提取 Knowledge，而不僅僅是從一個特定的中間層提取。這種跨層監督在訓練中是非常有用的。

下圖是 SemCKD model

首先是 Semantic Calibration Knowledge Distillation: 這裡顯示了 Teacher Network 和 Student Network，分別有目標層和學生層。對於 Student model 的每一層，都會有一個特定的特徵映射。這些特徵映射被投影到三個獨立的形式，以便與目標層的空間維度對齊。

再來是 Attention Allocation: 這裡顯示了 Key、Query 和維度降低與成對相似性以及特徵映射。學習的 Attention Allocation 可以幫助 Student model 專注於最具語義相關性的資訊，以實現有效的 Distillation。首先計算每個堆疊特徵映射之間的成對相似性，然後通過生成的 Query 和 Key vectors 之間的接近程度獲得 attention weights。



Practical Issue 2: Detecting the Novel Objects

1. (20%) CLIP[4] is a highly renowned large-scale vision-language model. It undergoes pretraining using a vast amount of paired text-image data through contrastive learning. CLIP[4] performs well on zero-shot classification task and has been the subject of further research by numerous researchers. Please describe the training process of CLIP and how it inference zero-shot classification.

train process:

首先是資料收集，CLIP 預訓練在大量的圖像與文字描述配對數據上，然後模型架構包含兩個主要組成，一個圖像編碼器和一個文本的編碼器。圖像編碼器通常是 CNN(ResNet)和 ViT，而文本編碼器是 Transformer 模型。在主要訓練的部分 CLIP 的核心理念是使用對比學習，在共享的嵌入空間中對齊圖像和文本的表徵。在訓練過程中，模型顯示一個 batch 配對的圖像和文本。目標是最大化正

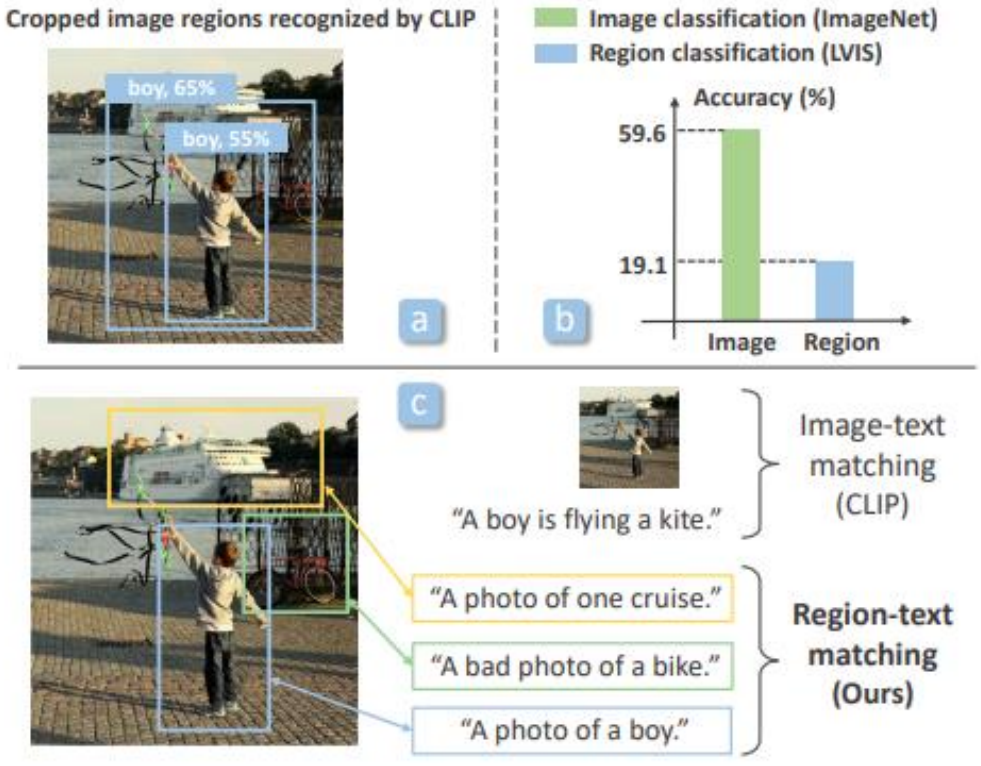
確配對之間的相似性，同時最小化錯誤配對之間的相似性，主要方法是透過 cosine similarity 讓匹配的圖文對相似度最大，不匹配的相似度最小，最後對調整大小後的圖片進行 Random Square Crop 來實現數據增強。

inference zero-shot classification:

首先為了進行 zero-shot classification，為數據集中的每個類別標籤創建文本提示。例如，如果類別是「貓」和「狗」，提示可能是「一張貓的照片」和「一張狗的照片」。接著使用文本編碼器對這些提示進行編碼，為每個類別創建文本嵌入，之後將待分類的圖像通過圖像編碼器生成圖像嵌入，再接著計算 cosine similarity，最後選擇與圖像嵌入相似性最高的類別文本嵌入作為圖像的預測類別。

2. (20%) In the Open-Vocabulary Object Detection task, it is intuitive to perform zero-shot classification using CLIP[4] after locating the unseen object. However, the paper RegionCLIP[5] points out that doing so presents certain challenges. Please describe these issues and explain how this paper addresses them.

首先，主要的問題是因為 Open-Vocabulary Object Detection(OVD)任務的目標是在訓練階段標記的基礎類別數量有限的情況下，實現對新類別的檢測。這些新類別由一個開放詞彙在推理階段時定義。換句話說，OVD 的核心思想是在 base class 的數據上進行訓練，然後完成對 unseen target 數據的識別和檢測，然而 CLIP 模型是被訓練用來匹配整個圖像與文本描述，並沒有捕捉到圖像區域和文本範圍之間的 fine-grained alignment。因此，如果直接應用這種模型來識別物體檢測的圖像區域會導致表現下降，所以為了解決這個問題，後來提出了一種叫 Region CLIP 的新方法，這個方法一定程度的擴展了 CLIP，使它能夠學習區域級別的視覺表示，像是特徵、模式等等，從而實現圖像區域和文本之間的 fine-grained alignment。這個方法利用 CLIP 模型將圖像區域與模板標題匹配，然後預訓練模型以在特徵空間中對齊這些區域-文本對
如下圖所示:


Cropped image regions recognized by CLIP
boy, 65%
boy, 55%
a
Image classification (ImageNet)
Region classification (LVIS)
Accuracy (%)
59.6
19.1
Image
Region
b
c
"A boy is flying a kite."
Image-text matching (CLIP)
"A photo of one cruise."
"A bad photo of a bike."
"A photo of a boy."
Region-text matching (Ours)