DLCV-Hw3 Report

R11942180 電信碩二 黃湛元

Problem 1: Grading - Report (15%)

1. Methods analysis (3%)

   - Previous methods (e.g. VGG and ResNet) are good at one task and one task only, and requires significant efforts to adapt to a new task. Please explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.

     CLIP（Contrastive Language–Image Pre-training）模型之所以能在各種圖像分類數據集上實現 zero-shot 性能，主要是因為它採用了一種與傳統的影像處理模型不同的訓練方式。CLIP 同時對大量的圖像和相對應的文本標籤進行學習，訓練過程中使用對比學習（contrastive learning）來匹配圖像和描述它們的自然語言文本。這使得 CLIP 模型不僅學習到視覺特徵，還學習到這些特徵和語言之間的對應關係。因此，當給予 CLIP 一個新的圖像和一組描述該圖像的文字（即使是從未見過的類別），CLIP能夠通過比較圖像特徵和文字特徵的對應關係來進行有效的預測。這種學習策略讓 CLIP 在沒有額外訓練的情況下，就能對新任務展現出強大的泛化能力。

2. Prompt-text analysis (6%)

   - Please compare and discuss the performances of your model with the following **three** prompt templates:

     1. “This is a photo of {object}” 0.7764
     2. “This is not a photo of {object}” 0.6812
     3. “No {object}, no score.” 0.5528

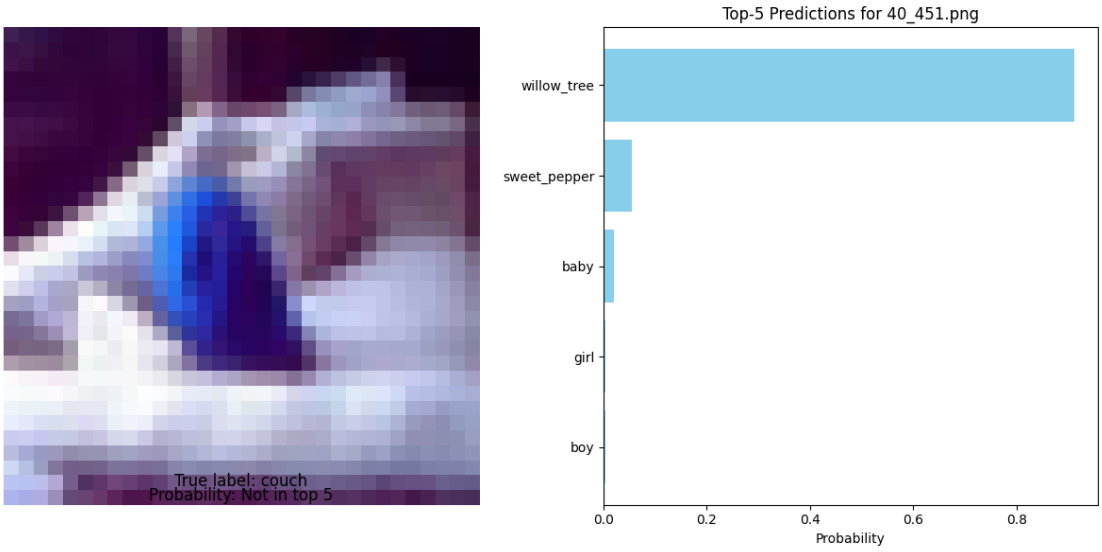
   - “This is a photo of {object}”：這個模板直接肯定了物體的存在，給出了一個明確的指示，模型能夠根據這個明確的指示找到對應的視覺特徵，因此得分相對較高。
   - “This is not a photo of {object}”：這個模板進行了否定表達，這要求模型不僅要識別圖像特徵，還要理解否定的語義，這增加了模型的識別難度，因此性能有所

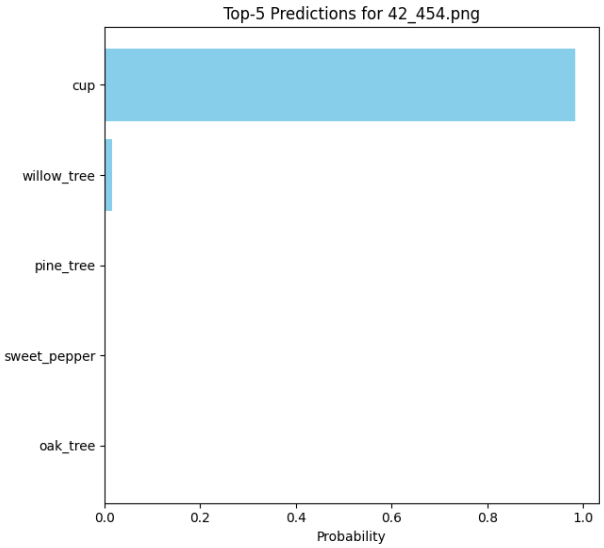
下降。

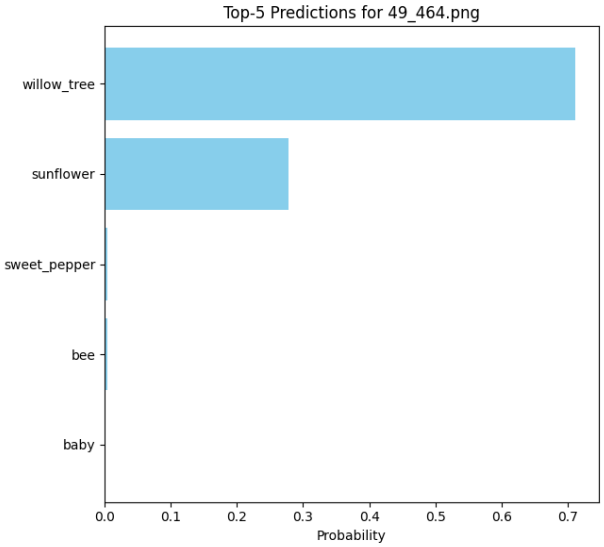
- “No {object}, no score.”：這個模板提供了一種條件性的表達，它將物體的存在與得分直接相連，這種表達方式對模型來說是一種更加抽象的理解，需要模型不僅識別圖像內容，還要理解複雜的語義關係，這是一個更高難度的任務，因此在三個模板中性能最低。

模板中的直接性和語義的複雜性對模型的性能影響顯著。更直接和簡單的語義提示有助於模型更好地識別和對應圖像內容，而複雜或抽象的表達則增加了識別的難度。

3. Quantitative analysis (6%)
   - Please sample **three** images from the validation dataset and then visualize the probability of the top-5 similarity scores as following example:







Problem 2: PEFT on Vision and Language Model for Image Captioning

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. (TA will reproduce this result) (2.5%)

   CIDEr: 0.8287734809295018 | CLIPScore: 0.7199272381786264

   我使用 CLIP ViT-L/14 做為我的 encoder，並在 decoder 加入 cross-attention block 和 adapter 來進行訓練。超參數如下：

```
ransform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    # 根據你的模型和訓練數據選擇適當的 mean 和 std。
    transforms.Normalize((0.485, 0.456, 0.406), (0.229, 0.224, 0.225))
])

batch_size = 64
n_epochs = 5
patience = 15
save_interval = 1  # 每 1 個 epoch 儲存一次模型
criterion = nn.CrossEntropyLoss(ignore_index=-100)
optimizer = torch.optim.AdamW(decoder.parameters(), lr=5e-4, weight_decay=1e-5)
scheduler = torch.optim.lr_scheduler.StepLR(optimizer, step_size=10, gamma=0.1)
```

2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

   Adapter: CIDEr: 0.8287734809295018 | CLIPScore: 0.7199272381786264

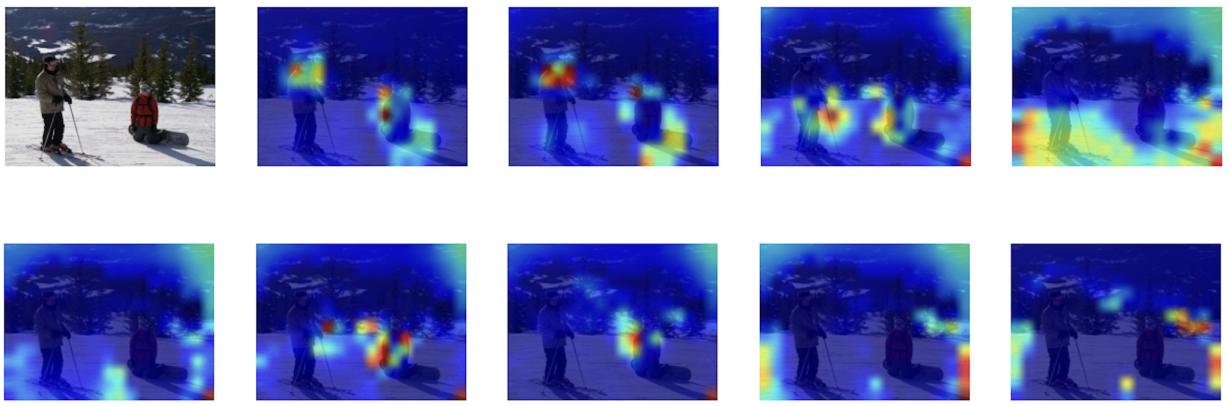
   Prefix:

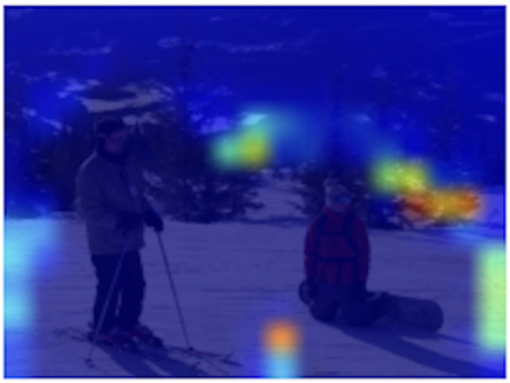
   Lora: CIDEr: 0.7947546419932882 | CLIPScore: 0.7143475877432399

3. TA will give you five test images ([p3_data/images/]), and please visualize the **predicted caption** and the corresponding series of **attention maps** in your report with the following template: (10%, each image for 2%)

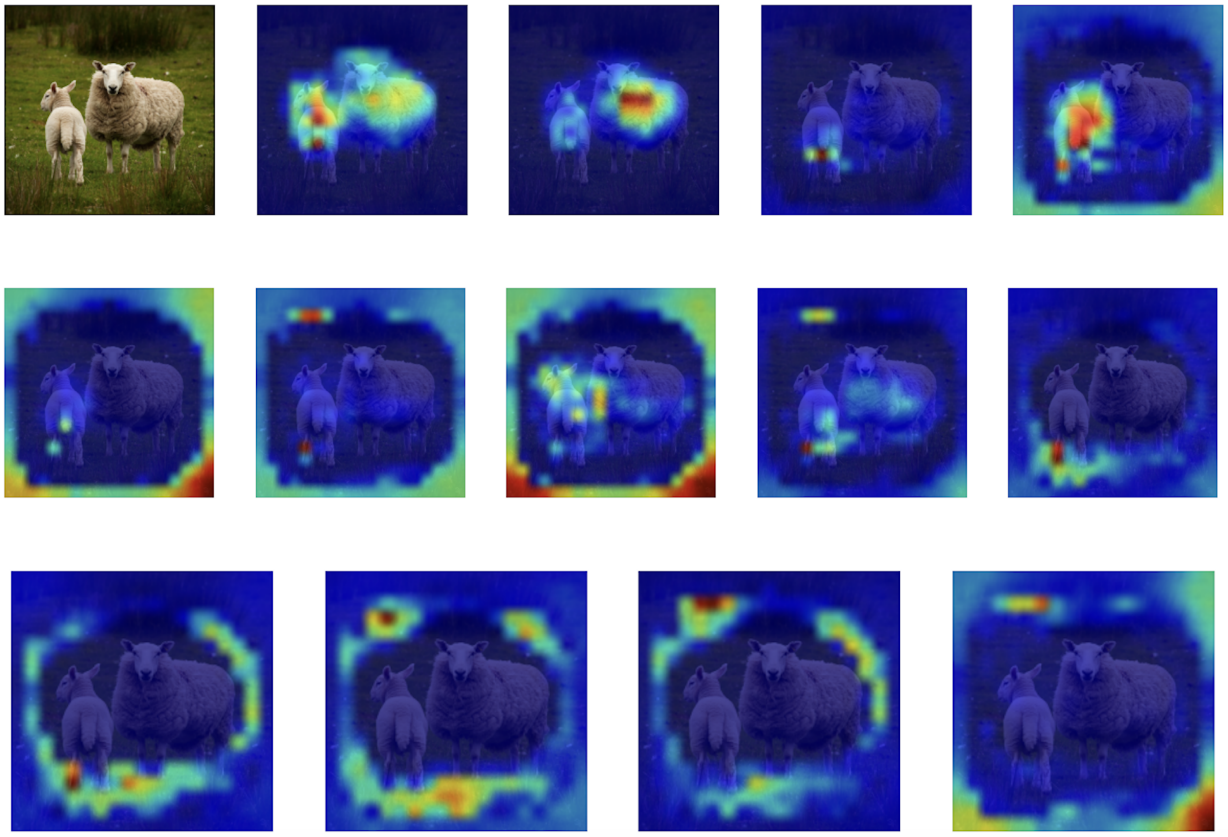
我選擇了最後一層 attention 做為視覺化的參考。

<endoftext> Two people on skis standing on a snowy hill <endoftext>

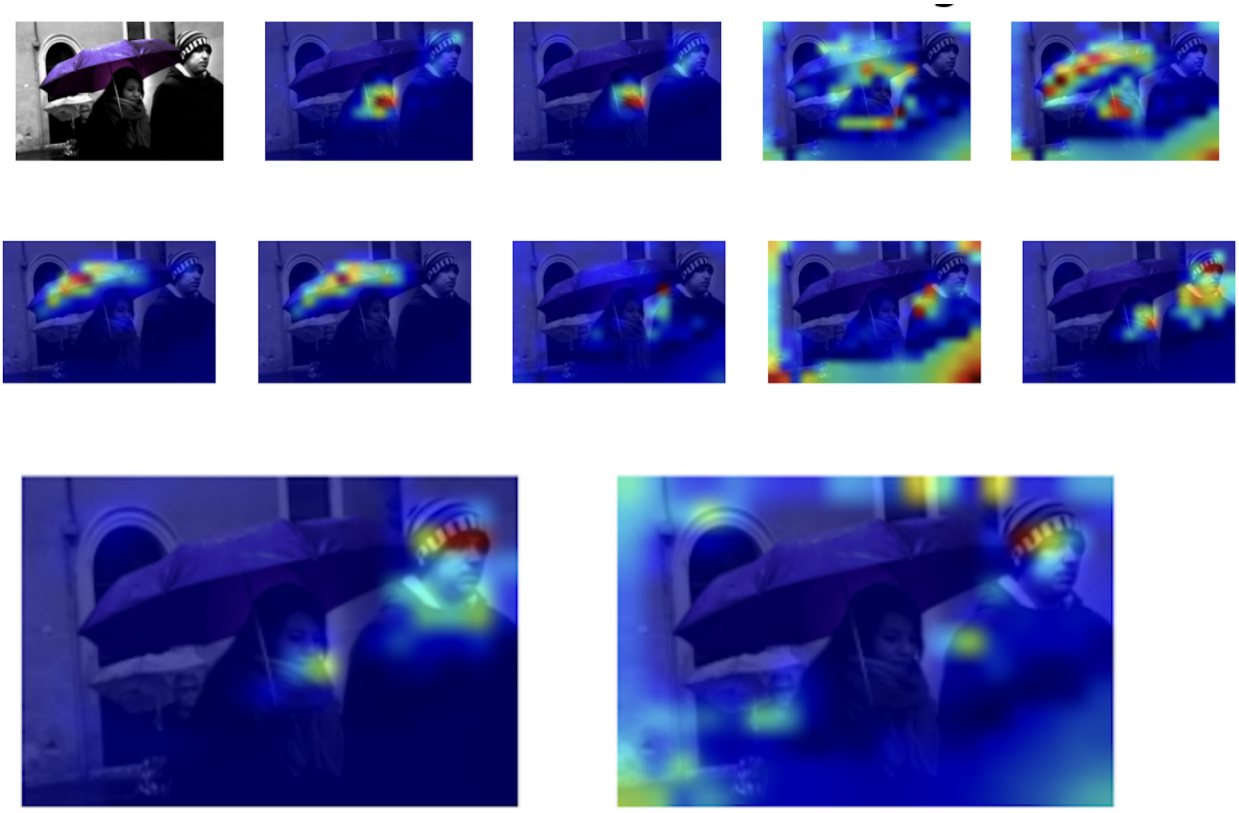




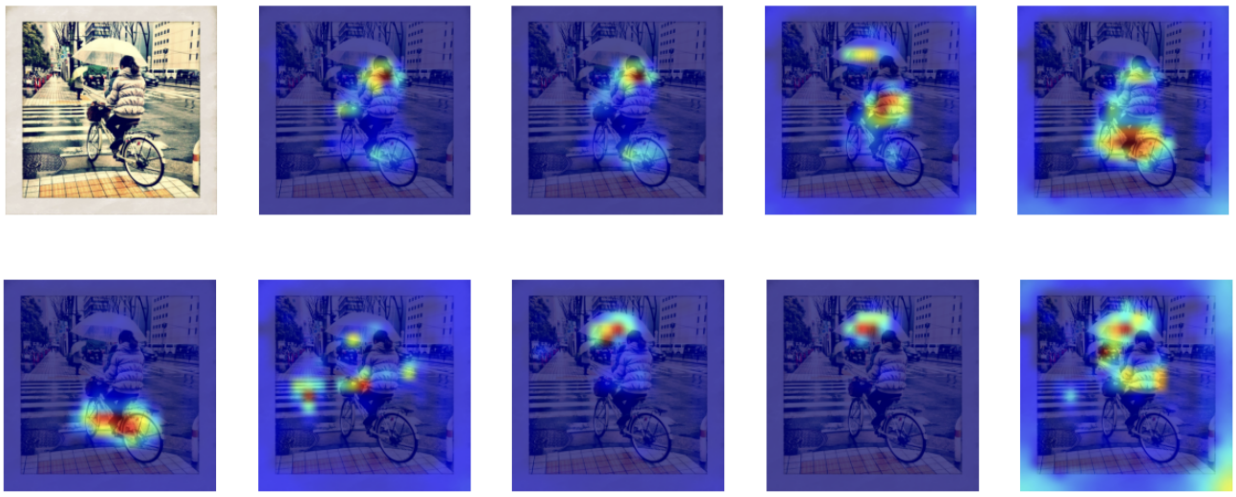
<endoftext> Two sheep standing next to each other on a lush green field <endoftext>



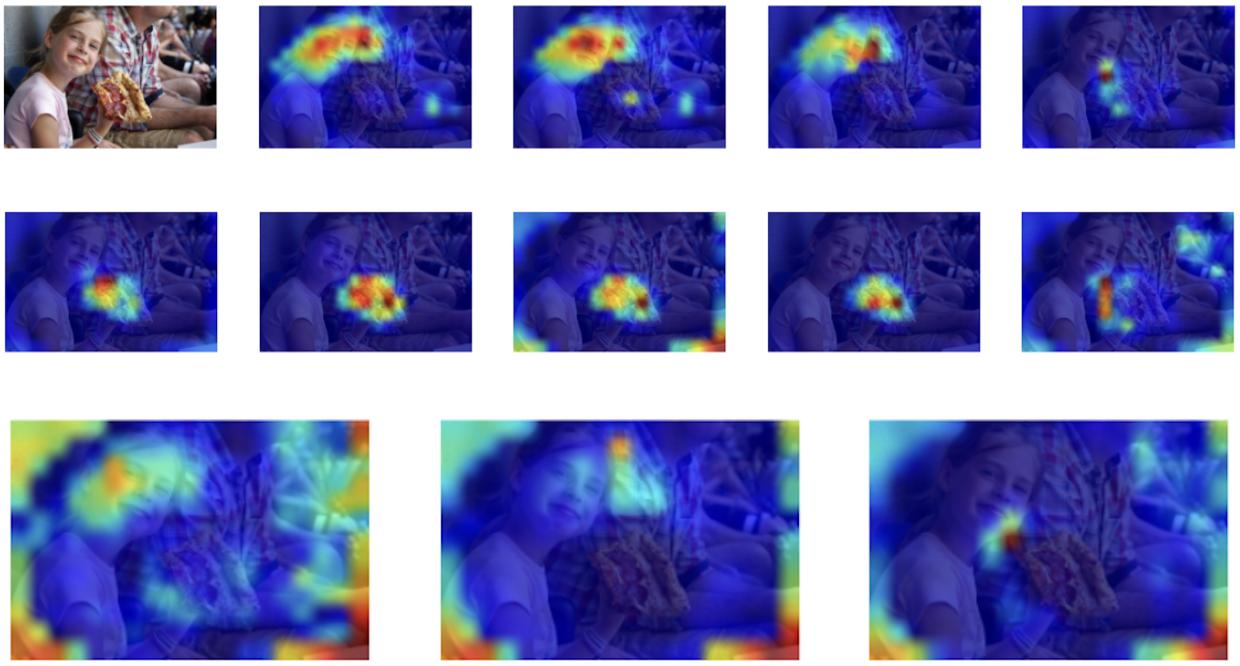
<endoftext> A woman holding a purple umbrella next to a man <endoftext>



<endoftext>A woman riding a bike with an umbrella <endoftext>



<endoftext> A young girl golding a slice of pizza in her hand <endoftext>



4. According to **CLIPScore**, you need to:
    a. visualize top-1 and last-1 image-caption pairs
    b. report its corresponding CLIPScore in the validation dataset of problem 2. (5%)
5. Analyze the predicted captions and the attention maps for each word according to the previous question. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption? (5%)
    a. **圖片1**：
        - 標題準確描述了場景，捕捉到主要元素：兩個人、滑雪板和雪地。注意力圖突出了每個單詞相關的區域，顯示出視覺焦點和描述性詞語之間的對應關係。
    b. **圖片2**：
        - 標題也似乎是對圖片的準確描述。注意力圖在提到羊時聚焦在羊身上，並在提到綠色草地時適當地擴展，顯示出單詞和視覺元素之間有很強的對應關係。
    c. **圖片3**：
        - 標題有效地描述了圖片的主要組成部分。注意力圖在提到女士和紫色雨傘時活躍，儘管並沒有強烈突出男士的存在，可能是因為標題更聚焦在女士和雨

傘上。

d. **圖片4**：
   - 標題正確識別了女士、自行車和雨傘。注意力圖在提到這些物體時照亮了對應的區域，表明模型的注意力與描述的物體一致。

e. **圖片5**：
   - 標題與圖片內容良好對齊，提到了年輕的女孩和 pizza。注意力圖在標題提到的相應時刻有效地聚焦在女孩的臉和 pizza 上。

結果看起來都還蠻合理的，反映了圖片的內容。注意力圖通常與標題中的特定單詞相對應，表明模型的注意力機制在生成描述性單詞時聚焦在圖片的正確區域。這表明模型已經學會將某些視覺模式與相關單詞聯系起來。

Reference

1. ChatGPT 幫忙撰寫程式碼 (Problem 1, 2)
2. Training / Validation/ Inference 相關的部分則參考了：
   HUNG-YI LEE (李宏毅) ML Source code
   https://speech.ee.ntu.edu.tw/~hylee/ml/2023-spring.php