DLCVHW3

學號：R12921053

系級：電機所碩一

姓名：周昱宏

Problem 1：Zero-shot Image Classification with CLIP

1. 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.

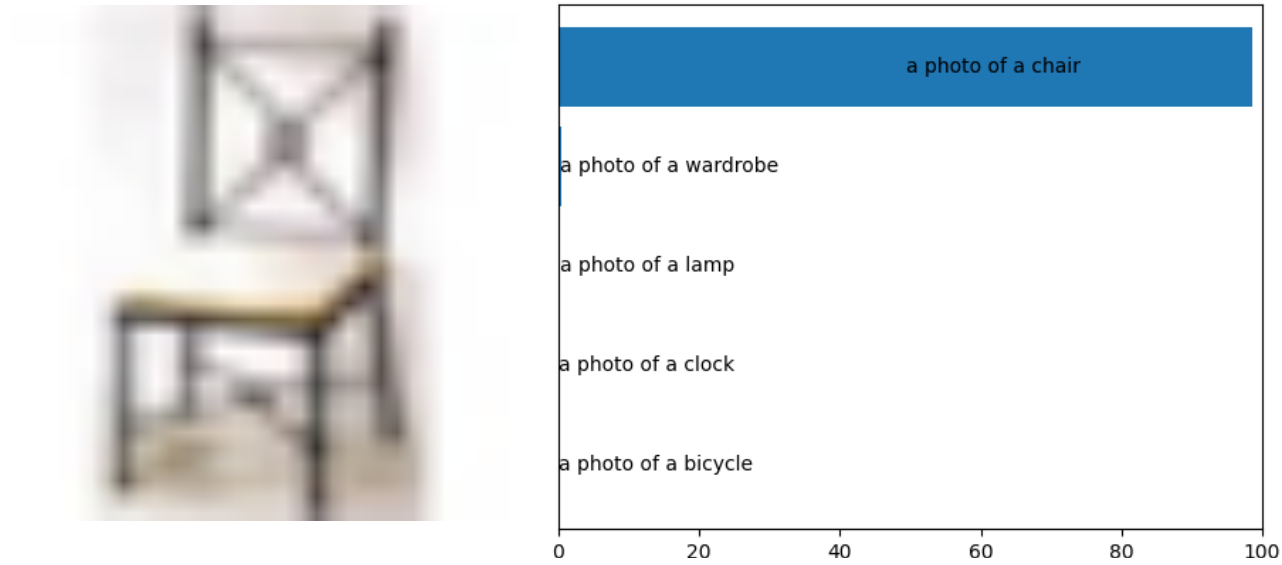
   我認為最主要的原因有兩個，首先是 CLIP 採用對比訓練的方式，讓模型學習如何讓相似的 sample 在 embedding 空間中的向量拉近，不相似的則推開，使得 CLIP 相較於 VGG 或 ResNet model 等傳統模型，CLIP 能夠更細節的捕捉圖像特徵，因為它不是只關心如何在特定資料集中找到最佳解函式，而是學習到不同 sample 的通用 representation。另外，CLIP 在預訓練階段使用「大量」現成的圖想與文字配對，其實變相讓 CLIP 有更多機會認識廣泛與多樣的 sample。

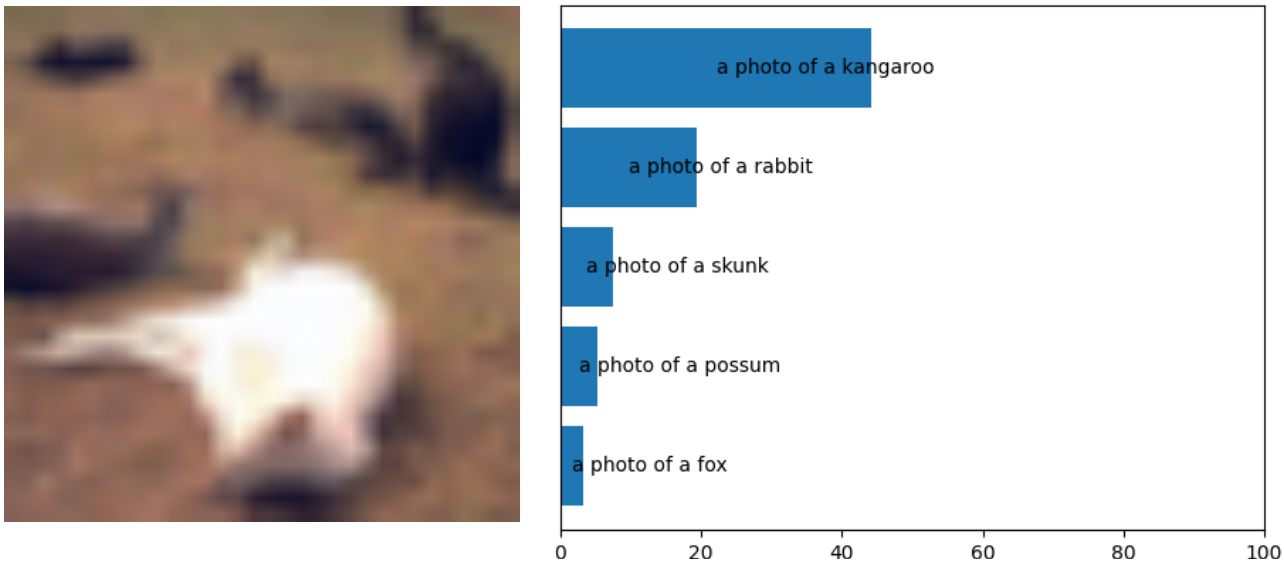
2. Please compare and discuss the performances of your model with the following three prompt templates/

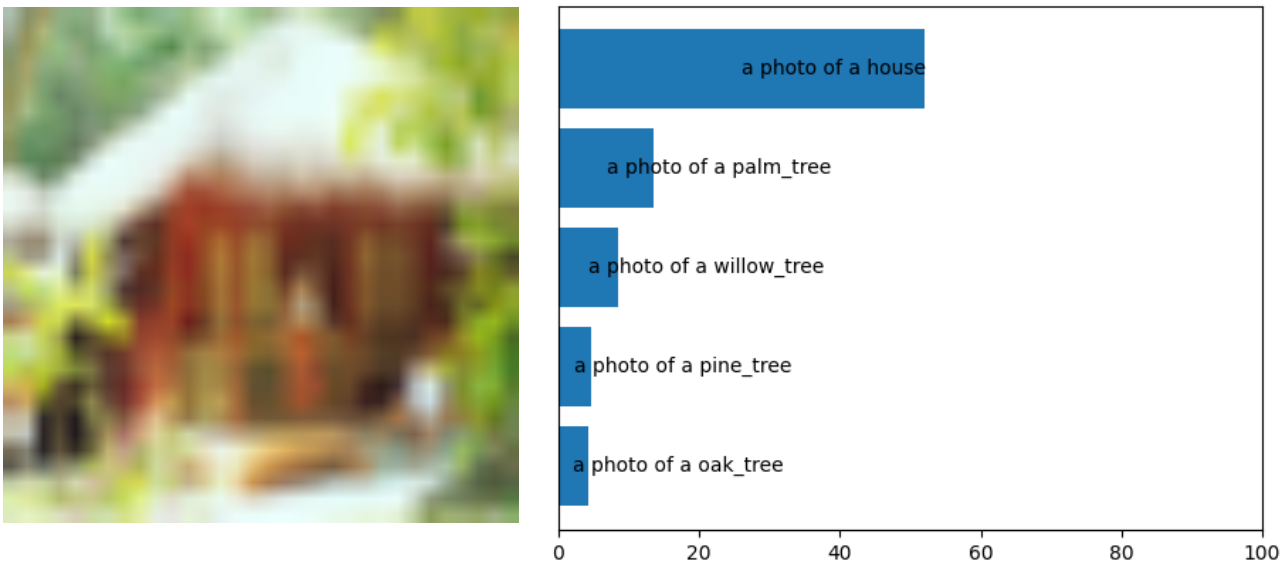
Prompt	Performance
“This is a photo of {object}”	0.6092
“This is not a photo of {object}”	0.6544
“No {object}, no score.”	0.5644

   我認為這三個 prompt 所體現的差別是在對模型訓練的期望，第一種 prompt “This is a photo of {object}”，主要是希望模型能夠辨識出圖像中特定物體，如此在訓練時需要接觸相當廣泛與多樣的物體，才能將文本中特定字詞與圖像中特定物體 match 起來。第二種 prompt “This is not a photo of {object}” 主要是訓練模型在反面情況下的預測能力，也就是讓模型正確判斷圖像中是否有缺少特定物體，如此相較於第一種 prompt 訓練目的會較為容易，因為任何無特定物體的圖像皆能與該特定物體配對。第三種 prompt “No {object}, no score.”，則是讓模型學習若圖像中沒有特定物體，則會給予較低的分數，如此不僅要求模型能夠識別特定物體，還需要讓模型了解特定物體與分數相關聯的能力，訓練上最為困難。

3. 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)

   CIDEr & CLIPScore :

   CIDEr: 0.8679049446813436 | CLIPScore: 0.7088603941404067

   Best Setting :

   - Training epoch = 7
   - Batch size = 14
   - Training steps = 64
   - Inference steps = 30
   - Learning rate = 1e-4
   - Weight decay = 1e-5
   - PEFT = adapter
   - Data augmentation : random rotation、color jitter、random horizontal flip

2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore.

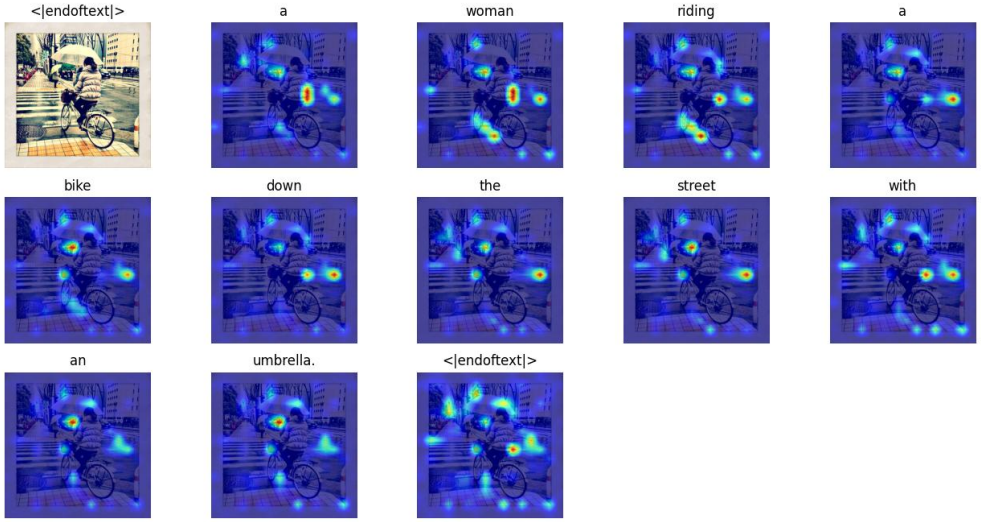
PEFT	CIDEr	CLIPScore
Adapter	0.867905	0.708860
Lora	0.850112	0.710086
Prefix tuning	0.830678	0.703865

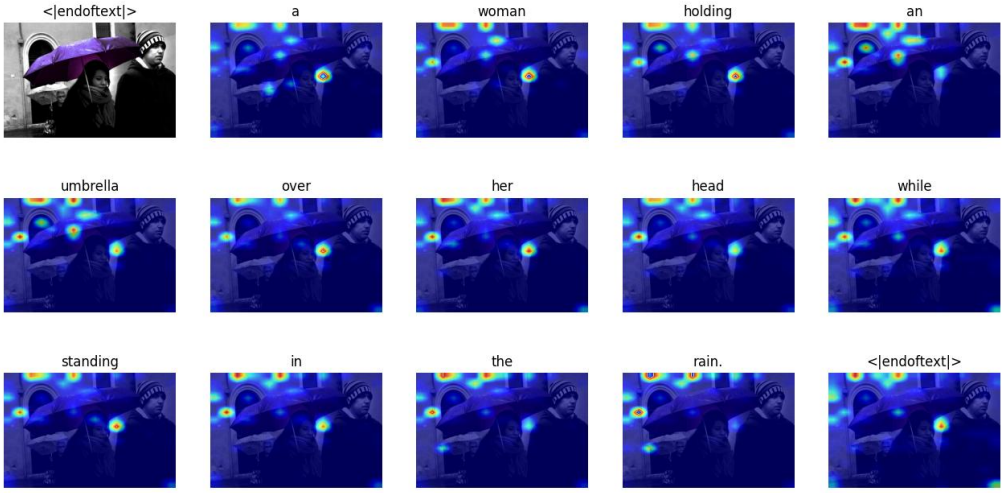
   Setting :

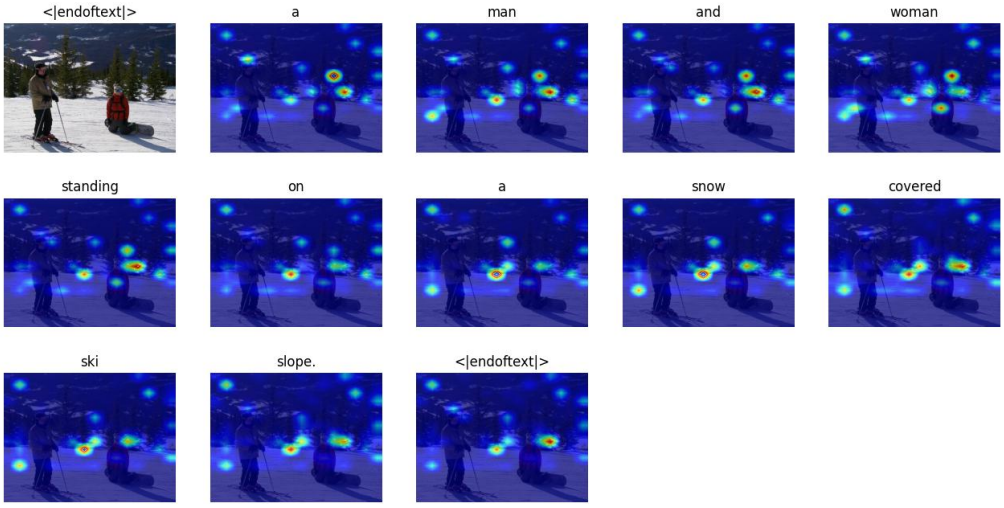
   - Choose the best score among 8 epochs
   - Batch size = 14
   - Training steps = 64
   - Inference steps = 30
   - Learning rate = 1e-4
   - Weight decay = 1e-5
   - Data augmentation : random rotation、color jitter、random horizontal flip

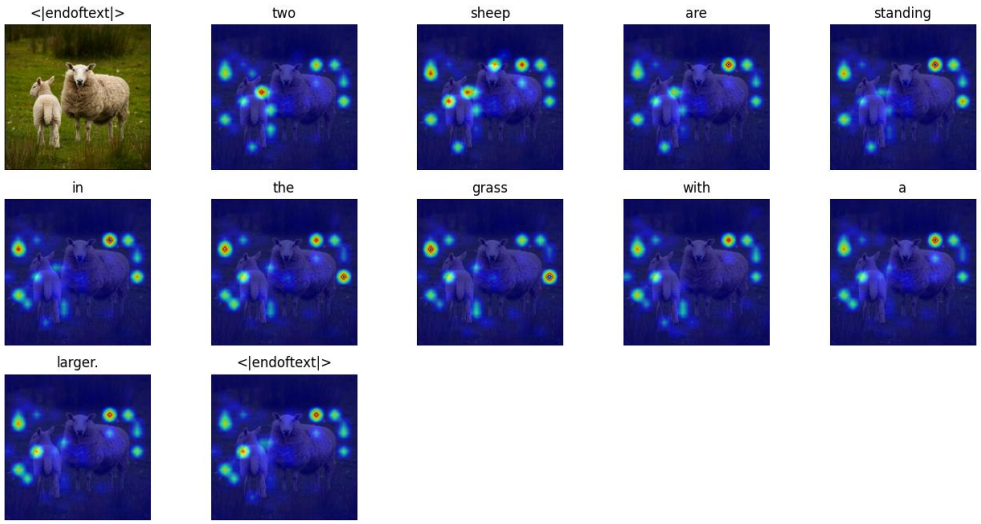
Visualization of Attention in Image Captioning

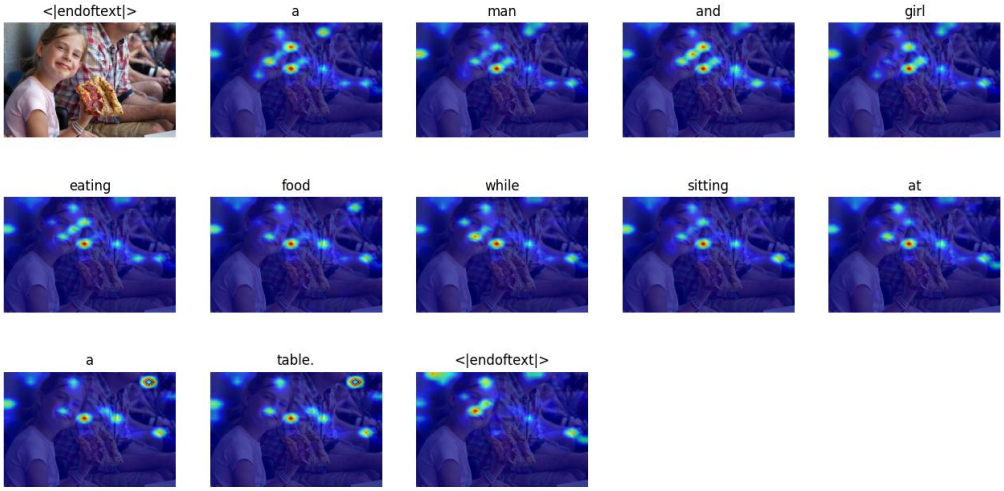
1. TA will give you five test images ([p3_data/images/]), and please visualize the predicted caption and the corresponding series of attention maps in your report with the following template







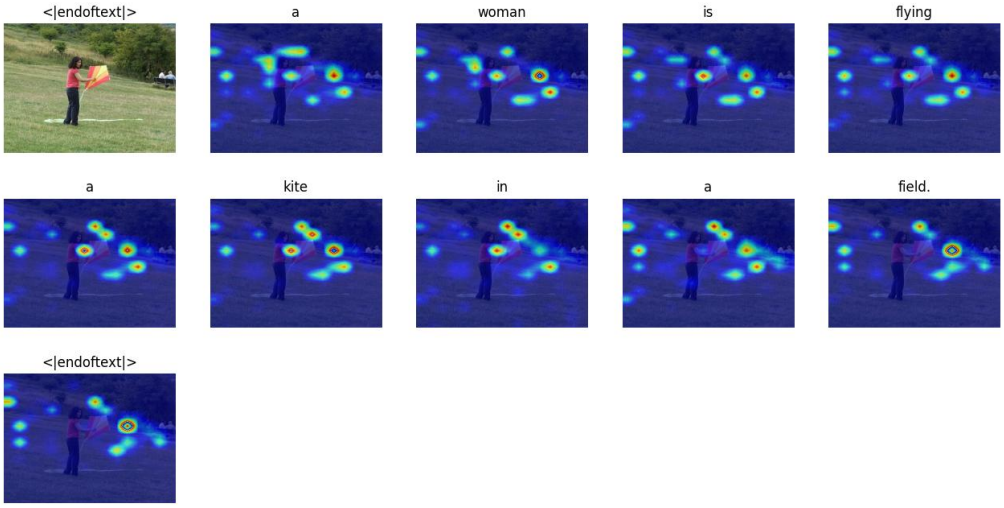

<|endoftext|>
two
sheep
are
standing
in
the
grass
with
a
larger.
<|endoftext|>


<|endoftext|>
a
man
and
girl
eating
food
while
sitting
at
a
table.
<|endoftext|>

2. According to CLIPScore, you need to : visualize top-1 and last-1 image-caption pairs and report its corresponding CLIPScore.

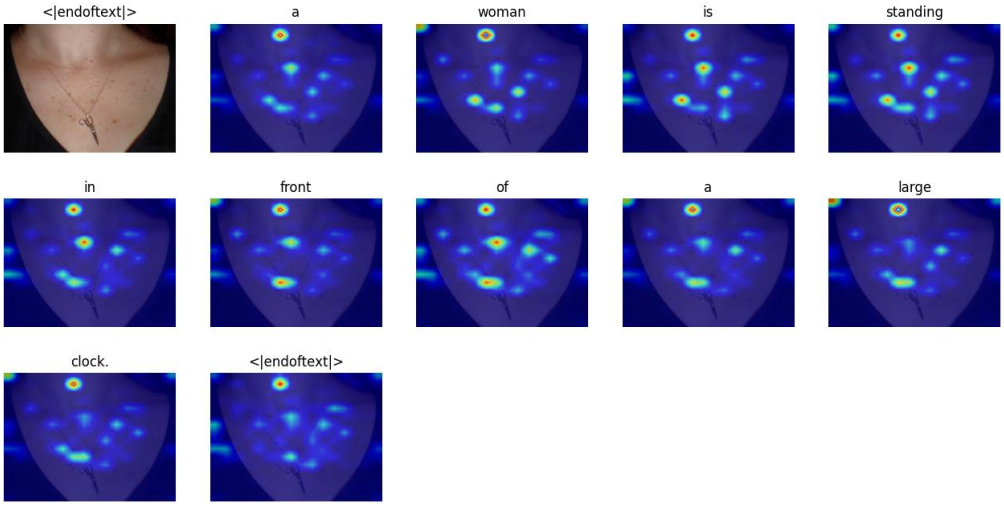
   Top-1 image-caption pairs : 000000179758.jpg, “a woman is flying a kite in a field.”

   with CLIPScore : 0.972290



   Last-1 image-caption pairs : 000000006393.jpg, “a woman is standing in front of a large clock.”

   with CLIPScore : 0.392151



3. 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?

   針對 top-1 的 image-caption pair 我認爲預測出來的文字相當符合圖片描述的結果，在 “woman” 的字眼中 attention 是偏左靠近女人一點，而 “flying a kite” 的 attention 則偏右邊一點較靠近風箏，最後 “in a field” 的 attention 朝兩邊發散，代表草地。而在 last-1 的 image-caption pair 明顯模型誤認為人脖子上的項鍊為時鐘，文字與圖片無法對應上，attention 也大致相同沒有特定位置的 highlight。

Reference

Problem 1

- p1_test.py：實作方式參考助教提供的 openai / CLIP 的 GitHub Repo (https://github.com/openai/CLIP) 修改的。
- p1_visualization：同 p1-test.py，視覺化部分使用 chatgpt 來輔助生成程式。

Problem 2

- p2_test.py：實作方式參考 saahiluppal / catr 的 GitHub Repo (https://github.com/saahiluppal/catr/tree/master?fbclid=IwAR1e3Npm4P3_ExycG2jLTX1DoXO_qaLMeYbA3zl551W7RTSkaMJbmlNLTJI)，包含 configuration 設置、evaluation 方式。而 beam search 部分則是參照「Transformers中的Beam Search高效實現」(https://blog.csdn.net/qq_27590277/article/details/107853325) 中的 code 來去做修改的，調整 decode 出來句子的最終選擇方式。
- p2_train.py：同 p2_test.py 參考 saahiluppal / catr 的 GitHub Repo，包含 configuration 設置 、training 方式。而 evaluation 方式則是參照助教提供的 p2_evaluation.py 中的函式呼叫。
- p2_dataset.py：同 p2_test.py 參考 saahiluppal / catr 的 GitHub Repo，包含 transform 方式、NestedTensor 相關函式、Resize 相關函式及 Dataset 相關函式建制方式。
- p2_evaluate_individual.py：實作方式參考助教提供的 p2_evaluation.py 來去修改的。
- p3_visualization.py：實作方式使用 chatgpt 來輔助生成程式。