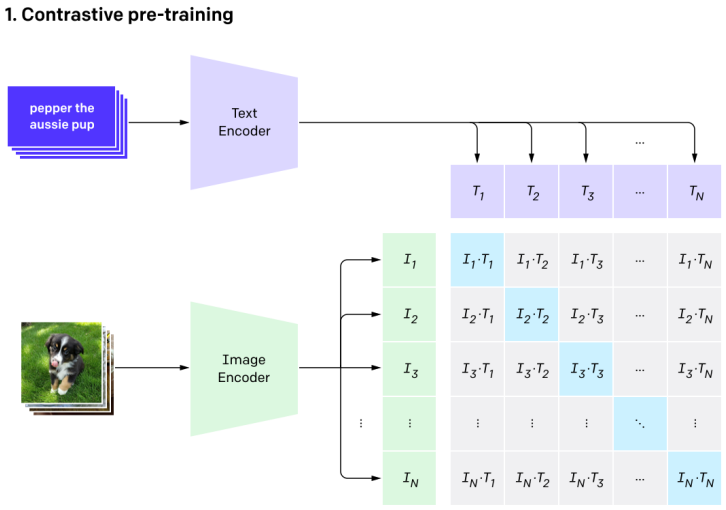
# DLCV HW3 Report

資工所 呂兆凱 R11922098

## Problem 1

1. Methods analysis

   OpenAI利用網路上大量的圖片與描述這些圖片的文字來訓練 CLIP 模型，透過盡量把圖片和文字映射到同一個空間的方式，讓模型直接學習文字與圖片之間的相關性。訓練方法則是採用 contrastive pre-training，在一個 batch 之中放入 N 張圖片與其相對應的描述文字，分別再放入 Image encoder 和 Text encoder ( 如下圖 )，而希望同張圖片的 Image 和 Text 的輸出越相似越好，不同圖片之間則相似度越小越好。因為 CLIP 在預訓練中並沒有針對特定的 dataset，且其學習的是圖片與文字的對應，這使 CLIP 模型可以更加通用在各領域，表現並不會侷限在某一 dataset 之上，並且也可以直接用在 zero-shot 的測試。




2. Prompt-text analysis
   a. “This is a photo of {object}”
      Accuracy: 1523/2500 (60.920%)
   b. “This is a {object} image.”
      Accuracy: 1709/2500 (68.360%)
   c. “No {object}, no score.”
      Accuracy: 1410/2500 (56.400%)

   可以看到 “No {object}, no score.” 的表現是全部最差的，可能是因為這句話並非是直接去描述圖片的句子，因此與 CLIP 模型預訓練時會使用描述此圖片的句子

的差距較大，所以使用這個 prompt 表現會較差。而前兩句都是有直接去描述此張圖片，因此效果會較接近原先所使用的 a photo of a {object}。

3. Quantitative analysis


**Problem 2**

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data.

    - Use pretrained ViT model as encoder : vit_large_patch16_224
    - Freeze encoder
    - Batch size : 100
    - Learning rate : 5e-5
    - DecoderLayer : nhead = 8
    - Decoder : num_layers = 6

    **CIDEr: 0.8124902289629274 | CLIPScore: 0.714995979663045**

2. Report other 3 different attempts (e.g. pretrain or not, model architecture, freezing layers, decoding strategy, etc.) and their corresponding CIDEr & CLIPScore.
    - **without pretrained ViT**
      CIDEr: 0.13139359904587458 | CLIPScore: 0.46620223427565155
    - **change to 8 decoder layers**
      CIDEr: 0.7989170064415443 | CLIPScore: 0.7194603482326893
    - **not freeze encoder**
      CIDEr: 0.7506919640784513 | CLIPScore: 0.6998040597488122

**Problem 3**

1. Visualize the predicted caption and the corresponding series of attention maps
    - bike.jpg

| <BOS> | a | woman | sitting | on |
|---|---|---|---|---|
|  |  |  |  |  |
| a | bicycle | in | a | city |
|  |  |  |  |  |
| street | with | a | woman | in |
|  |  |  |  |  |

| the | background | . | <EOS> |
|---|---|---|---|
|  |  |  |  |

- girl.jpg

| <BOS> | a | young | girl | sitting |
|---|---|---|---|---|
|  |  |  |  |  |

| on | a | bench | with | a |
|---|---|---|---|---|
|  |  |  |  |  |

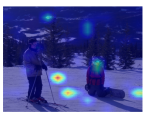
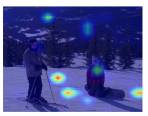
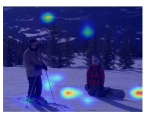
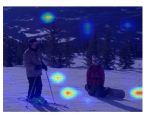
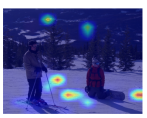
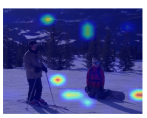
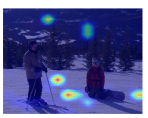
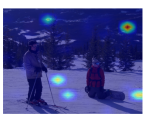
| large | slice | of | pizza | . |
|---|---|---|---|---|
|  |  |  |  |  |

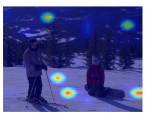
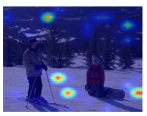
| <EOS> |
|---|
|  |

- sheep.jpg

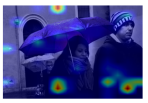
| <BOS> | a | sheep | and | a |
|---|---|---|---|---|
|  |  |  |  |  |
| baby | sheep | are | standing | in |
|  |  |  |  |  |
| a | field | . | <EOS> | |
|  |  |  |  | |

- ski.jpg

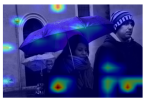
| <BOS> | two | people | on | skis |
|---|---|---|---|---|
|  |  |  |  |  |
| and | a | woman | in | the |
|  |  |  |  |  |

| snow | . | <EOS> |
|---|---|---|
|  |  |  |

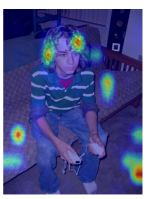
- umbrella.jpg

| <BOS> | a | man | with | an |
|---|---|---|---|---|
|  |  |  |  |  |

| umbrella | and | a | woman | with |
|---|---|---|---|---|
|  |  |  |  |  |

| a | umbrella | . | <EOS> |
|---|---|---|---|
|  |  |  |  |

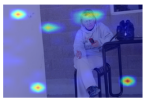
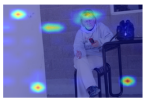
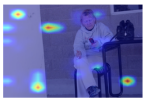
2. According to CLIPScore, you need to visualize in the validation dataset of problem 2.
    a. top-1 and last-1 image-caption pairs
        - top-1 :

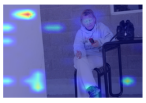
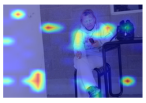
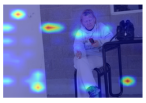
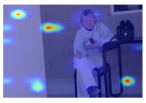
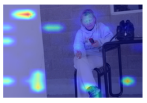
| <BOS> | a | young | boy | playing |
|---|---|---|---|---|
|  |  |  |  |  |

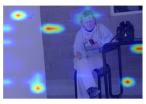
| a | game | of | nintendo | wii |
|---|---|---|---|---|
|  |  |  |  |  |

| . | <EOS> |
|---|---|
|  |  |

- last-1 :

| <BOS> | a | young | boy | wearing |
|---|---|---|---|---|
|  |  |  |  |  |

| a | white | shirt | and | a |
|---|---|---|---|---|
|  |  |  |  |  |

| tie | <EOS> |
|---|---|
|  |  |

b. its corresponding CLIPScore
    - top-1 : **1.016845703125**
    - last-1 : **0.37689208984375**

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?

   可以看到 top-1 圖片之中的 nintendo 的 attended region 集中在男生手上的手把上，以及 last-1 圖片中的 white 則是集中在圖片中女性的衣服上，而 last-1 圖片中的 boy 則是集中在圖片中的女性的短頭髮，而這個短頭髮使得模型誤判了圖中人物為男性，也使得其 CLIPScore 較低。
   因此我們也可以推斷 attended region 是會影響 corresponding word 的，而這些產生的 caption 也是合理的。

**Reference :**


- CLIP: Connecting Text and Images
- OpenAI 的 multimodal 神經網路 (下) CLIP: Connecting Text and Images
- SOURCE CODE FOR TORCH.NN.MODULES.TRANSFORMER
- Attention Map可视化