DLCV HW3 m11015q11 柯鐘奇

Problem 1.

1. a.

Model : pretrained vit model, use pretrained weight, unfreeze all layer.

Optimizer : SGD ( Learning rate = 1e-5, momentum = 0.9.

Scheduler : Step Learning Rate 0.5 per 10 epochs.

Architecture:

```
ViT(
  (patch_embedding): Conv2d(3, 768, kernel_size=(16, 16), stride=(16, 16))
  (positional_embedding): PositionalEmbedding1D()
  (transformer): Transformer(
    (blocks): ModuleList(
      (0): Block(
        (attn): MultiHeadedSelfAttention(
          (proj_q): Linear(in_features=768, out_features=768, bias=True)
          (proj_k): Linear(in_features=768, out_features=768, bias=True)
          (proj_v): Linear(in_features=768, out_features=768, bias=True)
          (drop): Dropout(p=0.1, inplace=False)
        )
        (proj): Linear(in_features=768, out_features=768, bias=True)
        (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
        (pwff): PositionWiseFeedForward(
          (fc1): Linear(in_features=768, out_features=3072, bias=True)
          (fc2): Linear(in_features=3072, out_features=768, bias=True)
        )
        (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
        (drop): Dropout(p=0.1, inplace=False)
      )
      (1): Block(
        (attn): MultiHeadedSelfAttention(
          (proj_q): Linear(in_features=768, out_features=768, bias=True)
          (proj_k): Linear(in_features=768, out_features=768, bias=True)
          (proj_v): Linear(in_features=768, out_features=768, bias=True)
          (drop): Dropout(p=0.1, inplace=False)
        )
        (proj): Linear(in_features=768, out_features=768, bias=True)
        (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
        (pwff): PositionWiseFeedForward(
          (fc1): Linear(in_features=768, out_features=3072, bias=True)
          (fc2): Linear(in_features=3072, out_features=768, bias=True)
        )
        (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
```

```
    (drop): Dropout(p=0.1, inplace=False)
  )
  (2): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (3): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (4): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
```

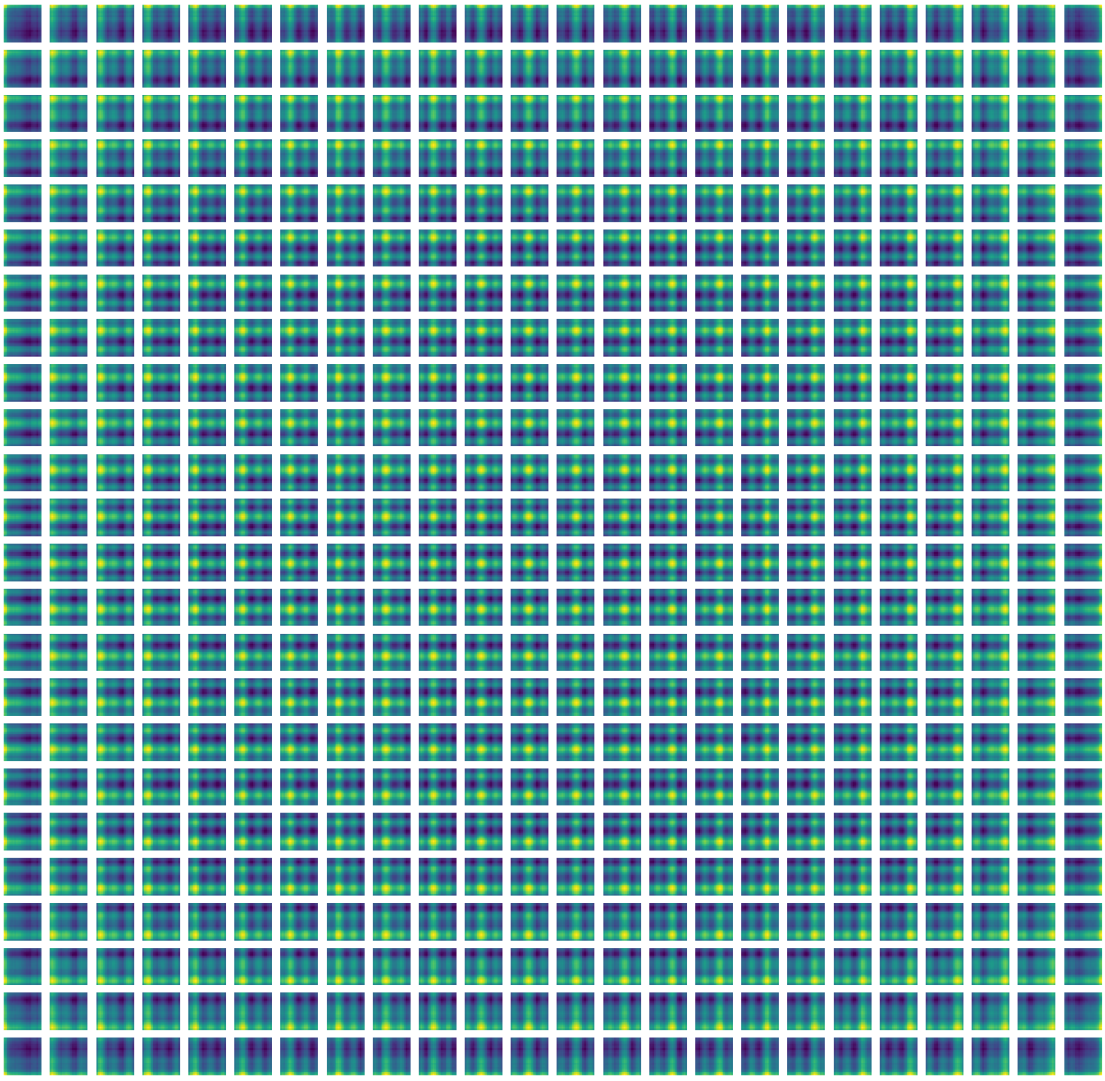
```
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (5): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (6): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (7): Block(
    (attn): MultiHeadedSelfAttention(
```

```
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (8): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
    )
    (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (drop): Dropout(p=0.1, inplace=False)
  )
  (9): Block(
    (attn): MultiHeadedSelfAttention(
      (proj_q): Linear(in_features=768, out_features=768, bias=True)
      (proj_k): Linear(in_features=768, out_features=768, bias=True)
      (proj_v): Linear(in_features=768, out_features=768, bias=True)
      (drop): Dropout(p=0.1, inplace=False)
    )
    (proj): Linear(in_features=768, out_features=768, bias=True)
    (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
    (pwff): PositionWiseFeedForward(
      (fc1): Linear(in_features=768, out_features=3072, bias=True)
      (fc2): Linear(in_features=3072, out_features=768, bias=True)
```

```
            )
            (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
            (drop): Dropout(p=0.1, inplace=False)
          )
          (10): Block(
            (attn): MultiHeadedSelfAttention(
              (proj_q): Linear(in_features=768, out_features=768, bias=True)
              (proj_k): Linear(in_features=768, out_features=768, bias=True)
              (proj_v): Linear(in_features=768, out_features=768, bias=True)
              (drop): Dropout(p=0.1, inplace=False)
            )
            (proj): Linear(in_features=768, out_features=768, bias=True)
            (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
            (pwff): PositionWiseFeedForward(
              (fc1): Linear(in_features=768, out_features=3072, bias=True)
              (fc2): Linear(in_features=3072, out_features=768, bias=True)
            )
            (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
            (drop): Dropout(p=0.1, inplace=False)
          )
          (11): Block(
            (attn): MultiHeadedSelfAttention(
              (proj_q): Linear(in_features=768, out_features=768, bias=True)
              (proj_k): Linear(in_features=768, out_features=768, bias=True)
              (proj_v): Linear(in_features=768, out_features=768, bias=True)
              (drop): Dropout(p=0.1, inplace=False)
            )
            (proj): Linear(in_features=768, out_features=768, bias=True)
            (norm1): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
            (pwff): PositionWiseFeedForward(
              (fc1): Linear(in_features=768, out_features=3072, bias=True)
              (fc2): Linear(in_features=3072, out_features=768, bias=True)
            )
            (norm2): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
            (drop): Dropout(p=0.1, inplace=False)
          )
        )
      )
      (norm): LayerNorm((768,), eps=1e-06, elementwise_affine=True)
      (fc): Linear(in_features=768, out_features=36, bias=True)
)
```

b. Accuracy:94.4%

2.



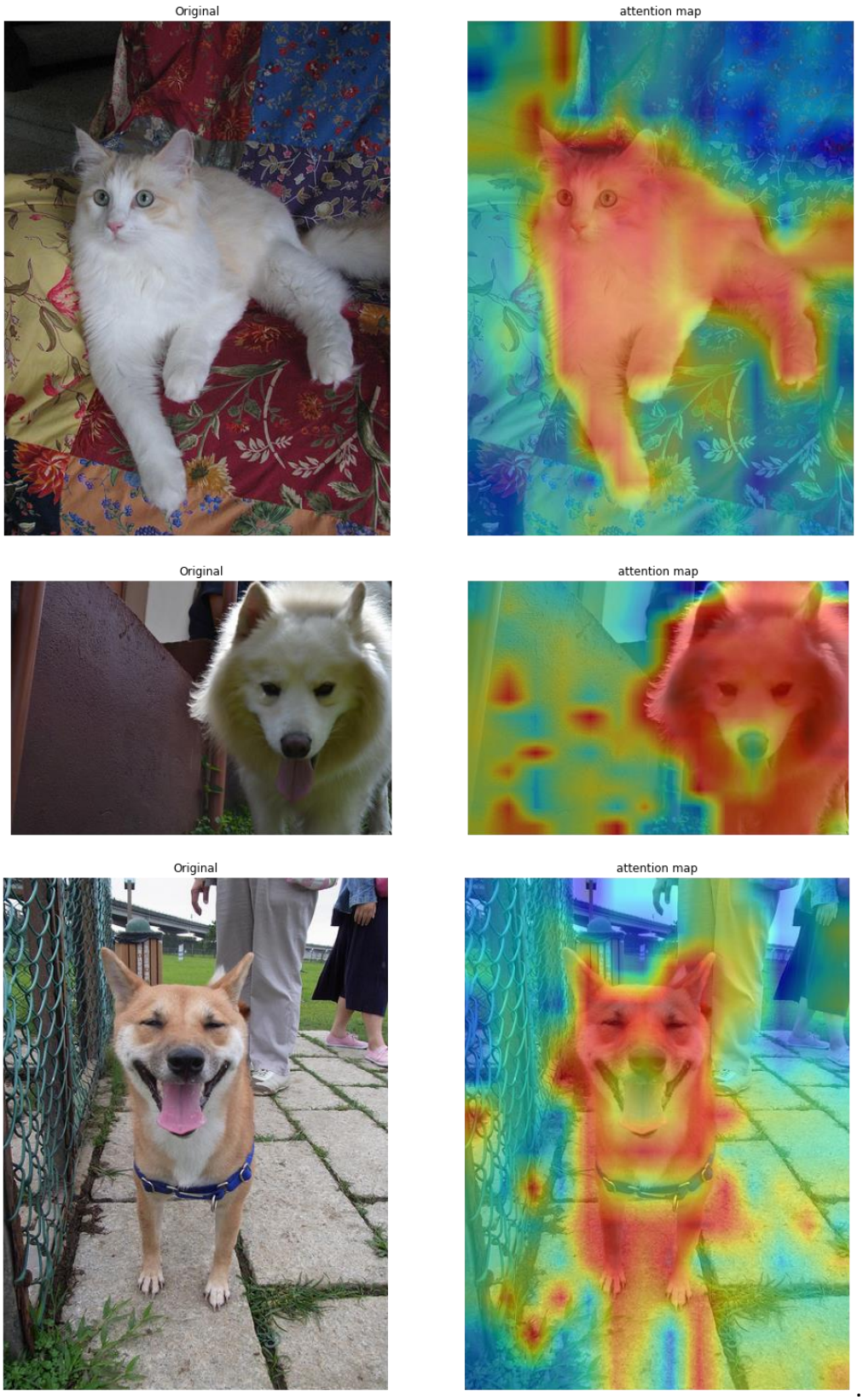
b.

取得方法：將 positional embedding 去掉 position 0 後，出來的結果 reshape 後減去 patch embedding，再用內積求得。

從上圖可知，position embedding 的 attention 大小會從自己向外擴散，

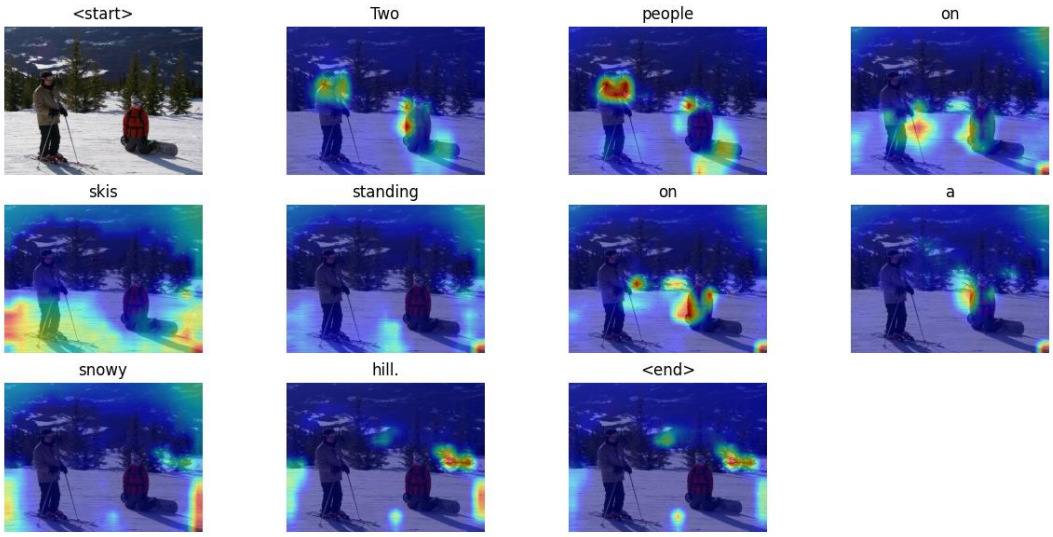
這也合理代表著位置訊息：越接近自己的區塊越需要被重視。

3. a.




b. 將 attention map 提取出來後，以 position 0(cls token)當作 query，attention 當 key 做矩陣相乘再做 head 平均。視覺化出結果後，可以發現當追蹤的物件和背景有明顯對比時(img1)，Attention map 的呈現就會有比較明顯的結果，當對比越來越小(img2, img3)，就很容易 attention 到其他地方。

Problem 2.




a.

Two 很不錯，有標示到兩個人。

People 都有標到人，但右邊那位稍微偏調。

On 表示有在滑雪板上。

Skis 似乎就有點偏掉了，影子的緣故導致滑雪板不好辨別。

Standing 完全跑掉，我認為是後面的背景導致。

On 在雪上也滿合理的。

A 單個人被標到。

snowy hill 就有點不合理了，範圍應該要大一點。

b.

學到了如何 trace 別人 model 的能力，也因為要改寫它得到我們要的 attention map ，需要更加了解它的架構以及流程，我認為最難的部分是找出 attention map 的輸出，要去爬 document 才會知道要的結果在哪裡。

Reference:

| | |
|---|---|
| Pretrained vit | https://github.com/lukemelas/PyTorch-Pretrained-ViT |
| Huggingface - vit | https://huggingface.co/docs/transformers/model_doc/vit |
| Position embedding | https://kazemnejad.com/blog/transformer_architecture_positional_encoding/ |
| github | https://github.com/google-research/vision_transformer/issues/55 |
| Attention map visualize | https://zhuanlan.zhihu.com/p/356798637 |
| MULTIHEADATTENTION | https://pytorch.org/docs/stable/generated/torch.nn.MultiheadAttention.html |
| 同學一 | 柯元豪 M11015Q02 |
| 同學二 | 黃柏翰 M11015Q12 |
| 同學三 | 易可鈞 M11015Q21 |