DLCV HW3

1

Pre-trained Weight: `openai/clip-vit-large-patch14`
Prompt-text: `"a photo of a "`
Accuracy: `81.365%`

1.1

Unlike its predecessors, CLIP is designed to understand and process both images and their corresponding text descriptions. This is a significant departure from traditional image classification models, which are typically trained on image data alone. The inclusion of text allows CLIP to develop a joint understanding of both visual content and the language used to describe it. This dual comprehension is crucial in its ability to generalize across various tasks without the need for task-specific training.

The training approach of CLIP also sets it apart. It employs a contrastive learning framework where the model is trained to match images with their corresponding text while also distinguishing them from non-matching pairs. This method doesn't just focus on recognizing specific visual patterns tied to certain labels but rather on understanding the broader context and associations between images and text. It helps the model develop a feature space that is robust and generalizable, enabling it to perform well across different types of image classification tasks.

Another factor contributing to CLIP's success in zero-shot performance is the scale and diversity of its training data. CLIP isn't limited to a specific domain or type of imagery; it's trained on a vast and varied dataset. This extensive and diverse training is crucial for the model to develop a broad understanding that can be applied to a wide range of images and scenarios. This is a significant advantage over models trained on more constrained datasets, which might excel in specific domains but struggle to adapt to new ones.

Lastly, CLIP's zero-shot learning capability is a hallmark of its design. Because it's trained on image-text pairs, CLIP can effectively perform classification tasks based on textual descriptions it has never encountered during its training. For example, given a description of an object or scene, CLIP can accurately identify images that fit this description, even if it hasn't been explicitly trained on a dataset containing those specific labels.

1.2

Prompt-text	“This is a photo of {object}”	“This is not a photo of {object}”	“No {object}, no score. ”
Accuracy	67.625%	69.485%	45.825%

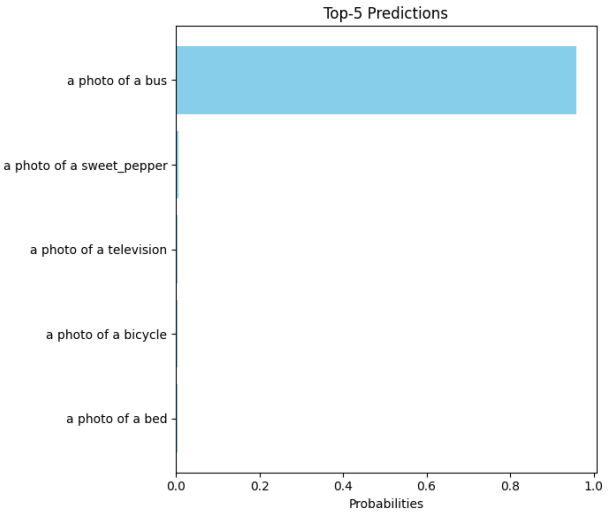
The first prompt follows a straightforward and direct description format. The model is tasked with affirming the presence of the specified object in the image. A 67.625% accuracy rate suggests that CLIP is fairly competent in identifying and confirming the presence of objects when the language used is direct and unambiguous. This is a typical use-case scenario for image recognition systems, demonstrating CLIP's strengths in basic recognition tasks.

the second prompt involves negation. I have taken the argmax to get the final prediction for each of the above three different prompt-texts, but interestingly this is a negative statement, and theoretically when taking the argmax, the accuracy of the prediction should be the lowest, however it is one of the highest scores in all the three prompt-texts, which means that the CLIP is not able to effectively understand whether the prompt-text is POSITIVE or NEGATIVE.

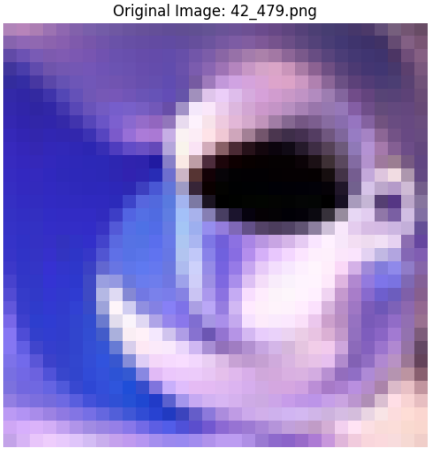
The third prompt shows a significant drop in accuracy. The phrasing here is more abstract and colloquial compared to the other two. It doesn't directly state the presence or absence of the object but rather implies it through a conditional statement. The lower accuracy could be due to several factors. First, the model may find it challenging to interpret the conditional and indirect nature of the statement. Second, such phrasing might be less common in the training data, making it a less familiar format for the model to process. This result highlights a potential area for improvement in CLIP, particularly in understanding more nuanced or non-literal language.

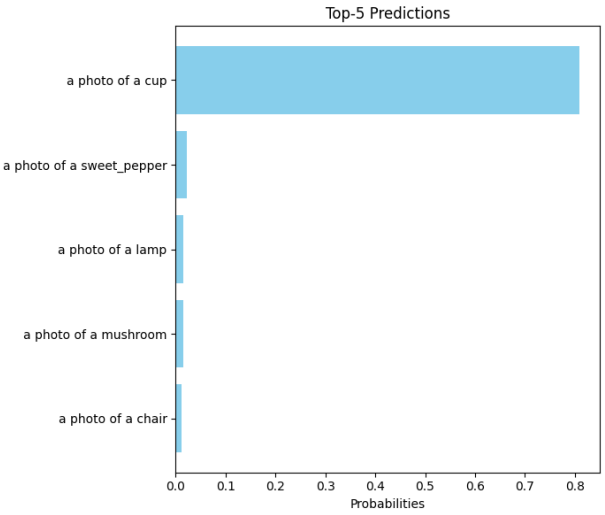
1.3

Original Image: 32_453.png

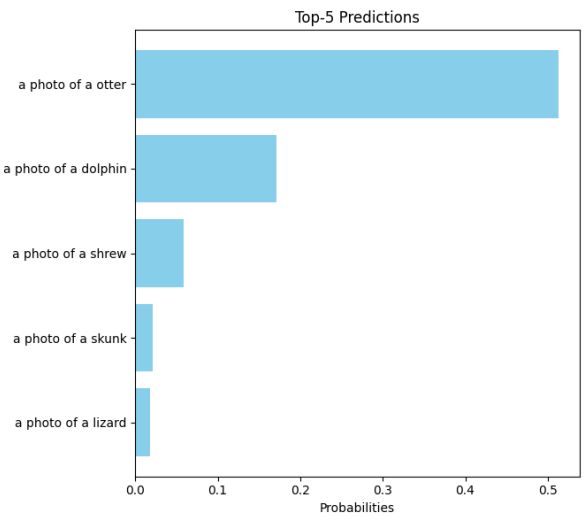


Original Image: 42_479.png





Original Image: 47_479.png



2.1.1

最佳的是CLIP(ViT-L/14@336px) + decoder(Adapter + Cross attention)

CIDEr	CLIPScore
0.9429	0.7168

當中只有Adapter + Cross attention是訓練的，

- cross attention的架構跟原來助教寫好的self attention基本一樣，只有將image feature輸入，並將k,v所學的權重由image feature提供。

- Adapter架構如下，把feature縮小至一半再減30，然後過activation fuction，然後放大回原來大小，過一次dropout，然後再把feature縮小至一半再減30，然後過activation fuction，然後放大回原來大小。

```
class Config:

    def __init__(self, checkpoint=None):
        self.n_layer = 12
        self.n_head = 12
        self.n_embd = 768
        self.vocab_size = 50257
        self.block_size = 1024
        self.checkpoint = checkpoint
```

```
class Adapter(nn.Module):
    def __init__(self, cfg) -> None:
        super().__init__()
        self.down_project = nn.Linear(cfg.n_embd, (cfg.n_embd // 2)-30)
        self.up_project = nn.Linear((cfg.n_embd // 2)-30, cfg.n_embd)
        self.activation = nn.GELU(approximate="tanh")
        self.dropout = nn.Dropout(0.1)

    def forward(self, x):
        x = x + self.up_project(self.activation(self.down_project(x)))
        x = self.dropout(x)
        x = x + self.up_project(self.activation(self.down_project(x)))
        return x
```

即完成一個adapter block，此block被放在cross attention block後面、mlp中間。

2.1.2

以下三種架構的訓練參數都一樣：

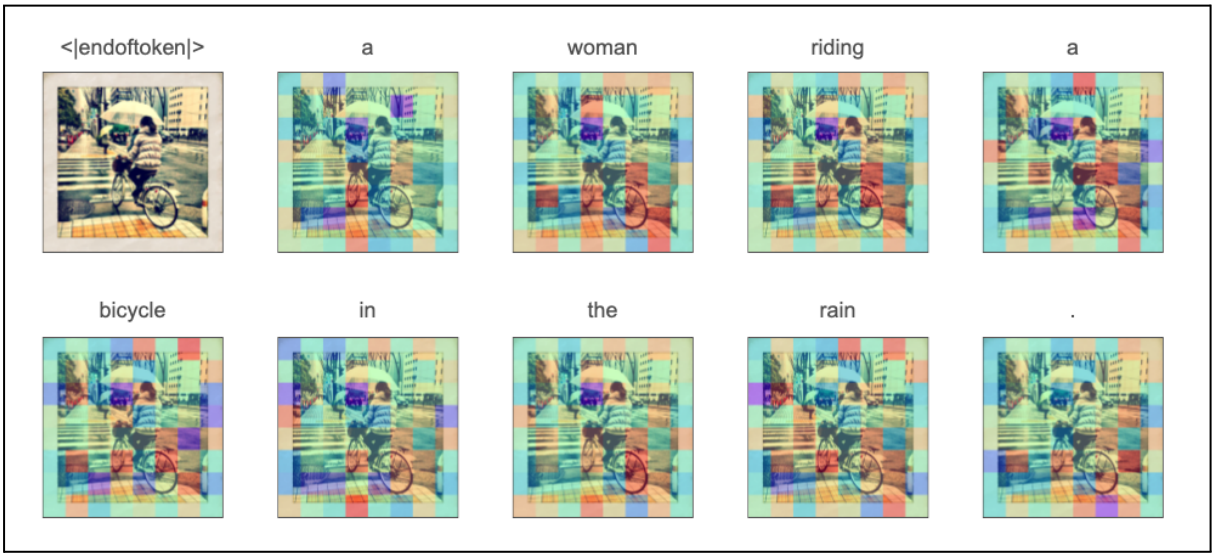
```
epochs = 10
batch_size=1
train_transform = transforms.Compose([
        transforms.ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2),
        transforms.RandomHorizontalFlip(),
        transforms.RandomRotation(15),
    ])
optimizer = torch.optim.AdamW(decoder.parameters(), lr=1e-4, weight_decay=1e-5)
scheduler = StepLR(optimizer, step_size=len(train_dataloader)/2, gamma=0.8)
```

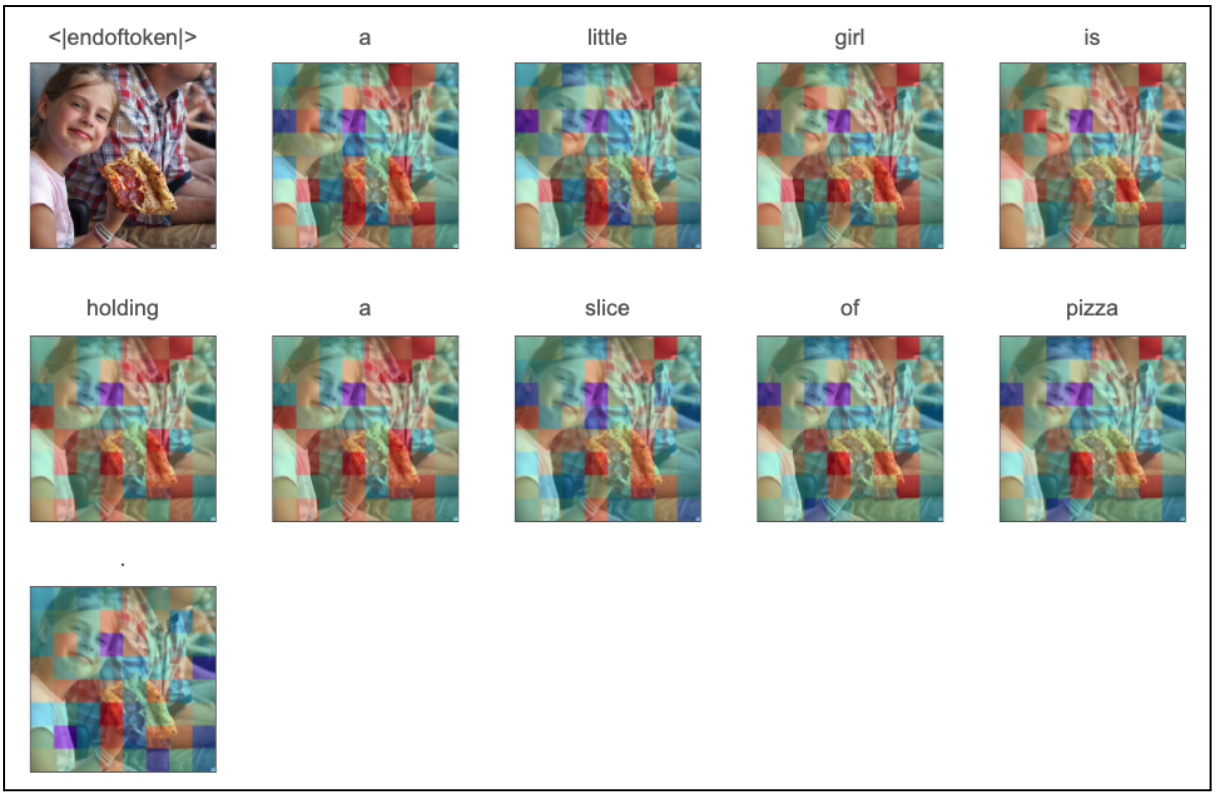
但我做了額外的training data前處理，把'\n'、句點前／後的空格，句子前空格等都刪除。

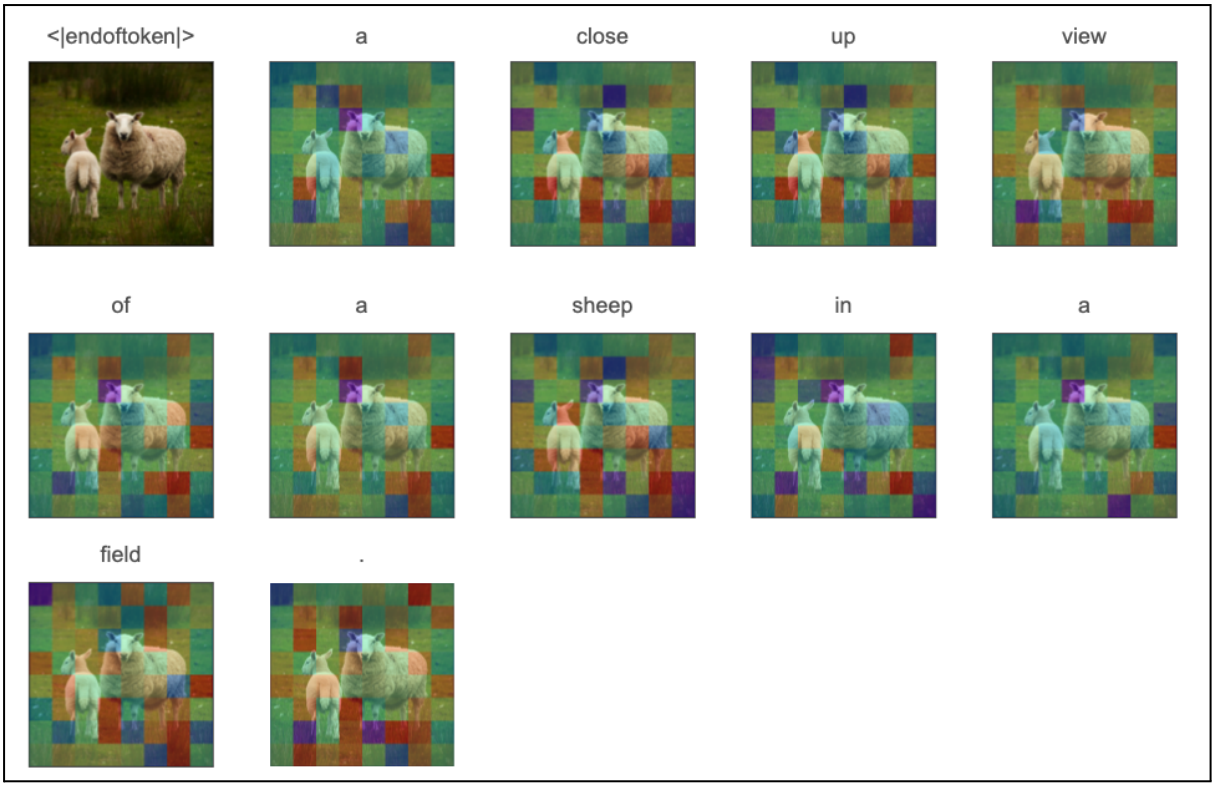
	CIDEr	CLIPScore
Adapter	0.943	0.717
Prefix tuning	0.745	0.688
LoRA	0.843	0.696

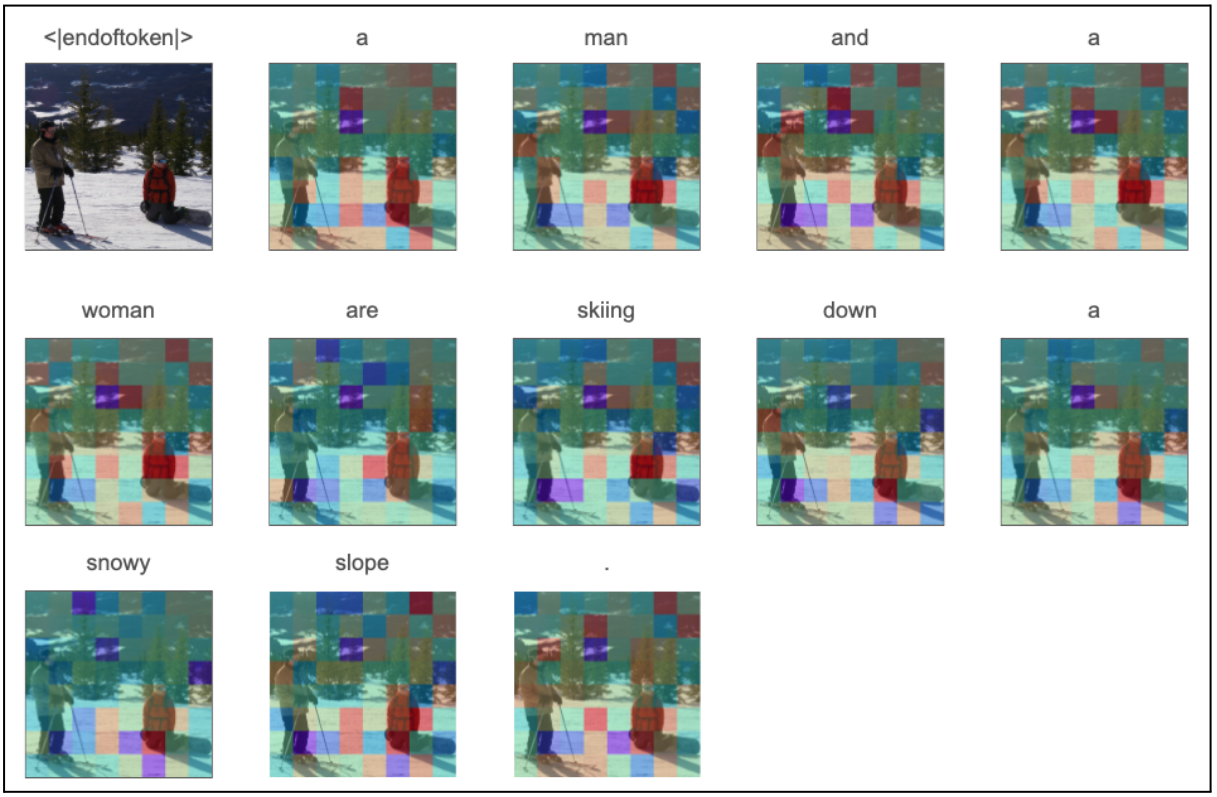
- Adapter在2.1.1已介紹過。
- Prefix tuning的cross attention部分與2.1.1相同，prefix的部分為size為(20, n_embd)，相當於加了20個token在prefix中，可是效果最差，我猜測是沒有訓練到模型內部太多東西所致。
- LoRA的cross attention部分與2.1.1相同，然後把attention裡的nn.linear都換成lora.linear，只訓練lora.linear，也有相當不錯的結果，應是lora.linear設計的降維方法能保留大部分精確度所致。

2.2.1


<|endoftoken|>
a
woman
riding
a
bicycle
in
the
rain
.

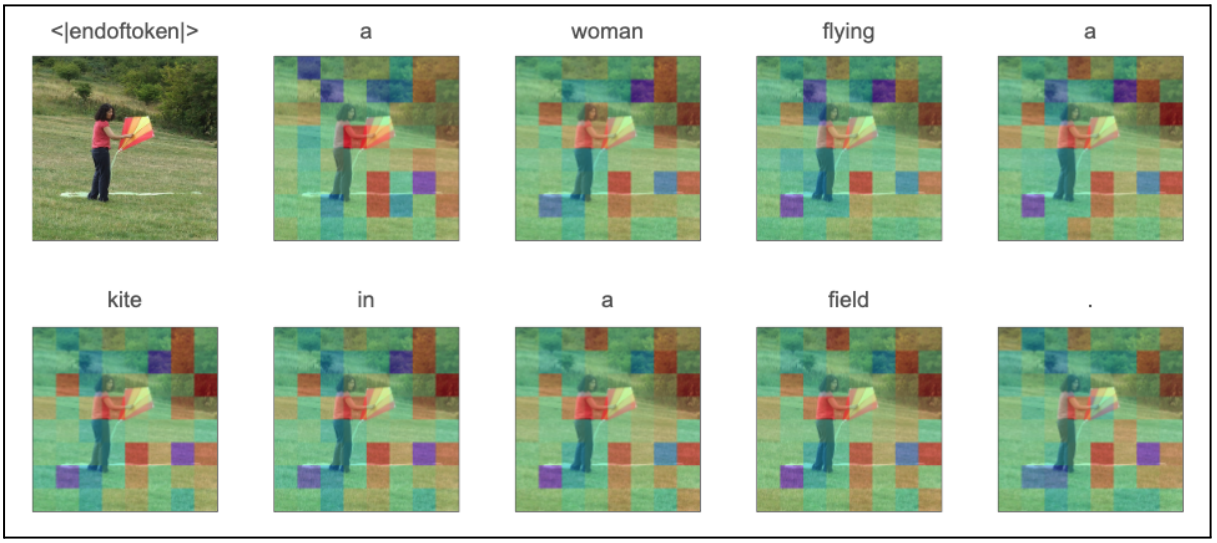

<|endoftoken|>
a
little
girl
is
holding
a
slice
of
pizza
.


<|endoftoken|>
a
close
up
view
of
a
sheep
in
a
field
.

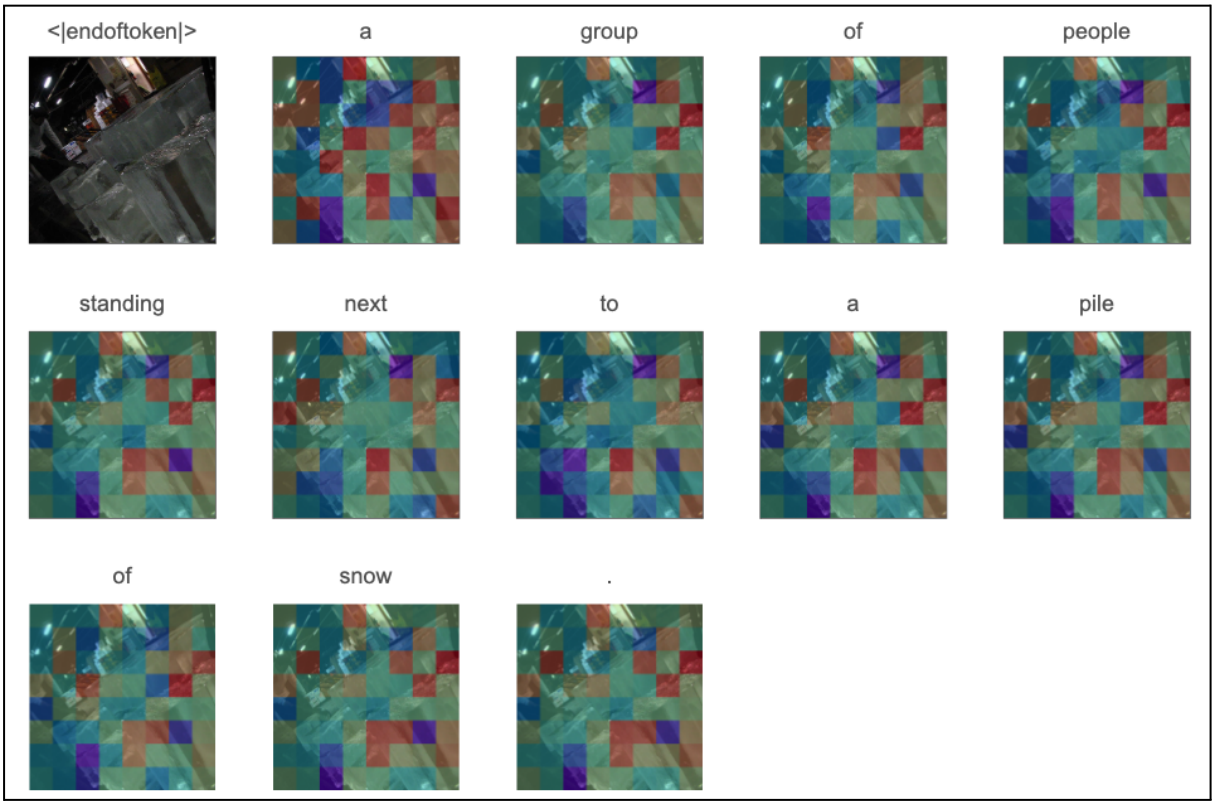

<|endoftoken|>
a
man
and
a
woman
are
skiing
down
a
snowy
slope
.

2.2.2

highest: 0.996



lowest: 0.435



2.2.3

可以看到highest的caption相當合理，的確就是一個woman在field中放風爭，但沒有看到明顯的word和圖片有配對到。而lowest的caption不合理，圖中也沒有看到任何人或雪，caption也沒有配對到。