1-1
以往的cnn模型是以物體照片及對應的classl去進行訓練，因此只能對已知的label去進行分類，但clip是以文字敘述作為照片的label，所以只要訓練良好的model即可對沒看過得物體進行分類。
1-2

*“This is a photo of {object}”*

```
accuracy 0.608

real    0m17.697s
```

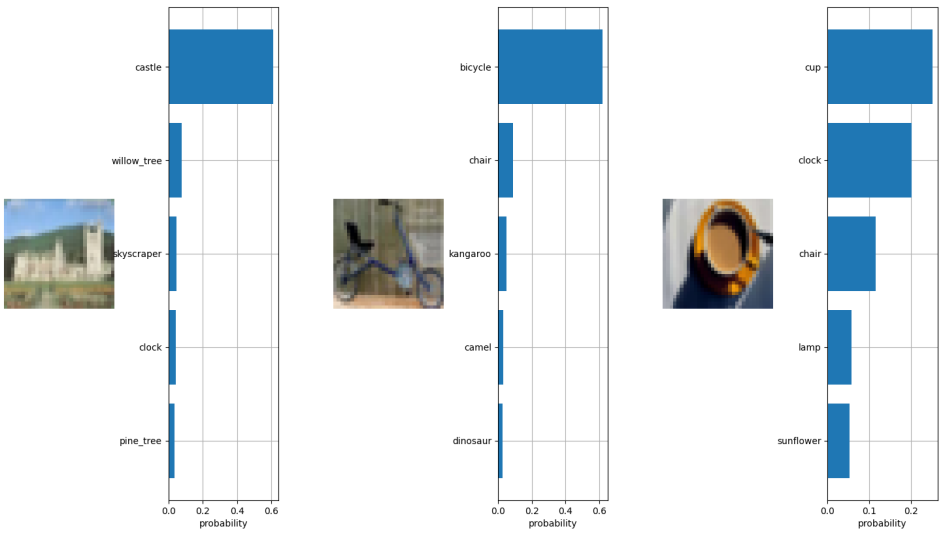
*“This is a {object} image.”*

```
accuracy 0.6816

real    0m21.355s
```

*“No {object}, no score.”*

```
accuracy 0.5628

real    0m17.801s
```

*1-3*



*2-1*

*encoder layer:5*

*decoder layer:4*

*backbone=resnet101*

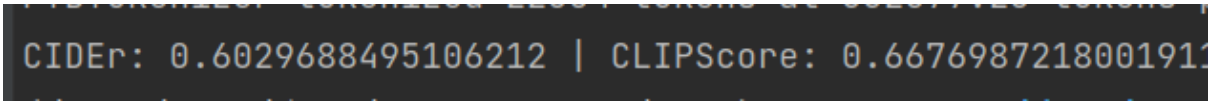
*#multi-head's head =8*

*dropout=0.1*


*2-2*

*encoder layer:6*

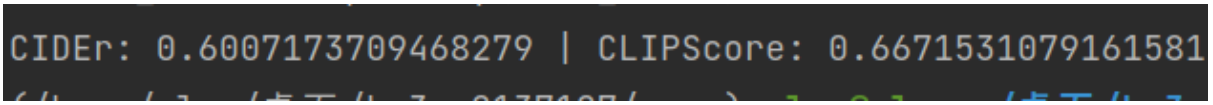
*decoder layer:6*




*encoder layer:6*

*decoder layer:6*

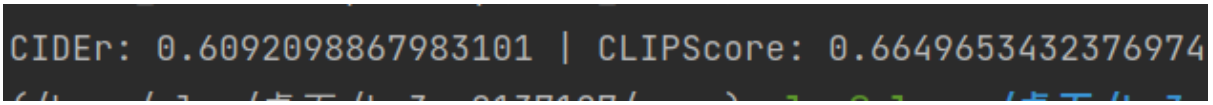
*dropout=0.2*



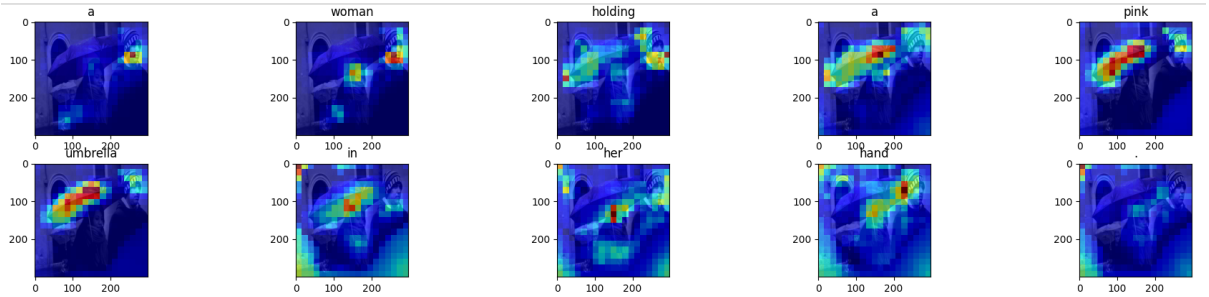

*backbone=resnet50*

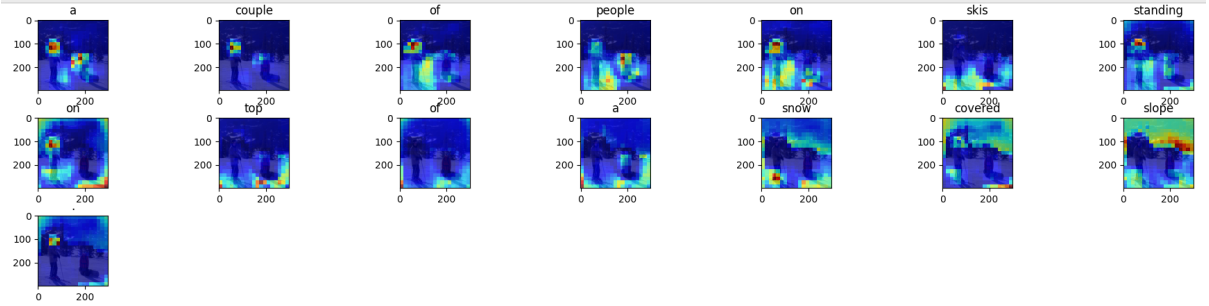
*encoder layer:4*

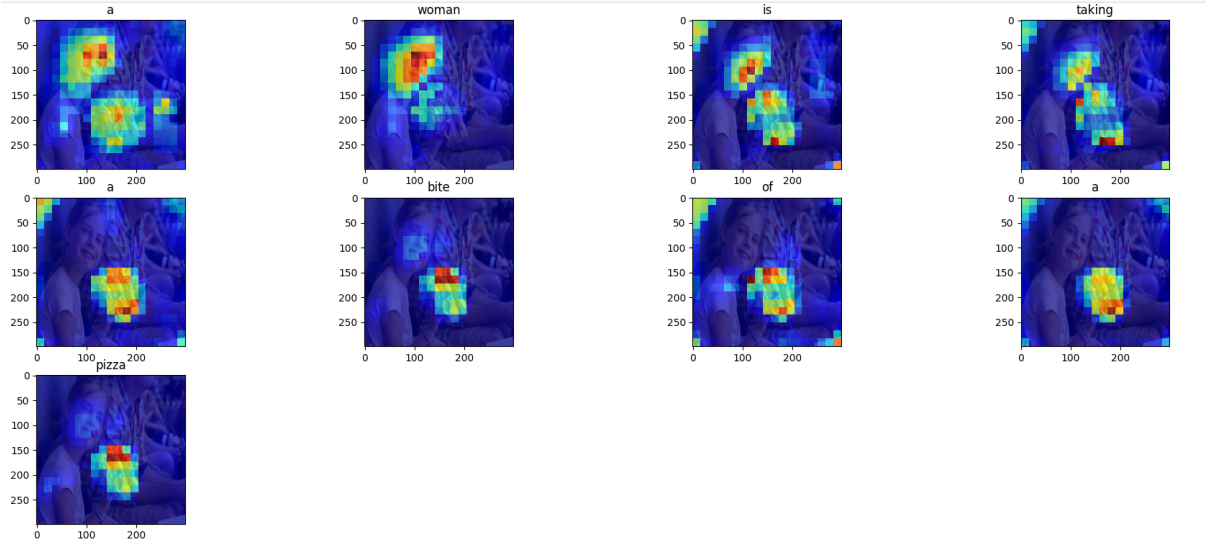
*decoder layer:4*

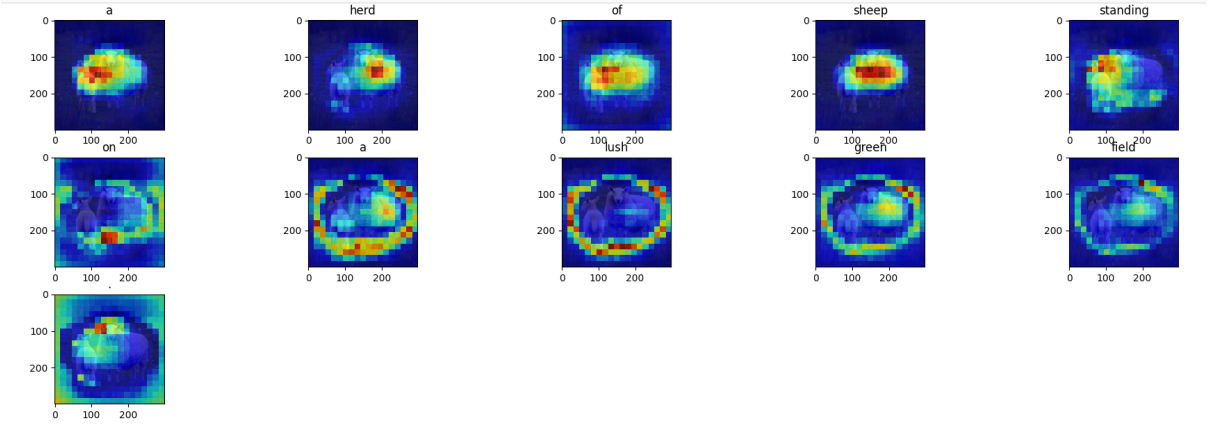


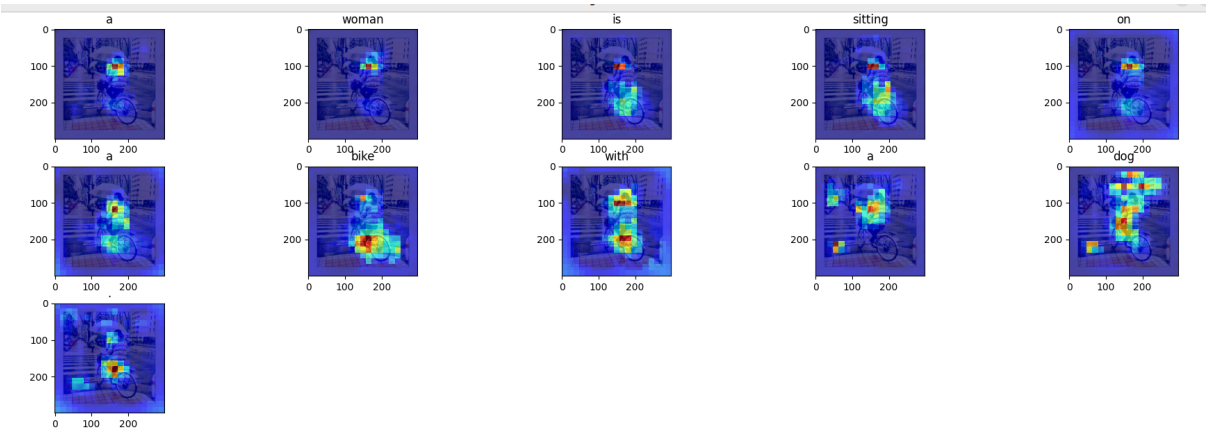

*3-1*


a
woman
holding
a
pink
umbrella
in
her
hand

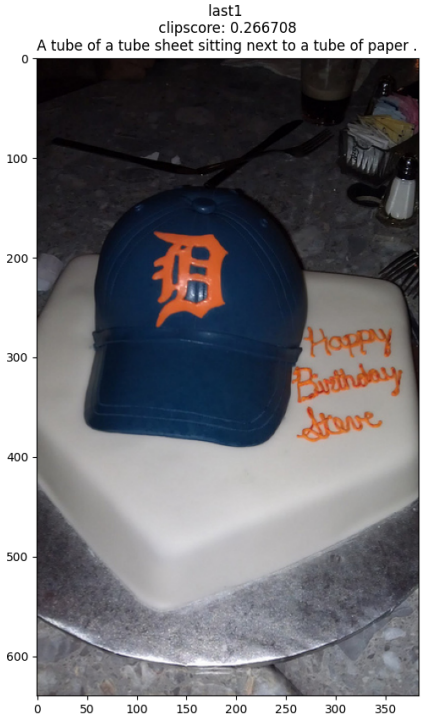

a
couple
of
people
on
skis
standing
on
top
of
a
snow
covered
slope


a
woman
is
taking
a
bite
of
a
pizza


a
herd
of
sheep
standing
on
a
lush
green
field




*3-2.*


*3-3*

就多數的照片如果它講的關鍵字是正確的，那它的確有正確的框出對應的目標，如第一組照片的雨傘或是粉色。第二組基本上大致正確，”兩個人”、”在滑雪板上”、”覆蓋雪的山坡”。第三組的”披薩”及”女人”也標示正確，但在介係詞等就比較沒道理。第四組的”群羊”及”草地”也都標示正確。第五組”女人”及”腳踏車”也正確，但最後說跟著狗但標示的卻是人及部份背景，這就明顯錯誤。就我的訓練結果看來，雖然分數沒有過*baseline*但依據這五張照片的描述及對應的位置大致上是正確的，只是對於介係詞或是非特定物體的詞彙則有些無法標示正確，而在最後一組的狗的部份為較明顯的錯誤。